Altérations génétiques et épigénétiques dans la leucémie myélomonocytaire chronique - Modulation par les agents déméthylants

Jane Merlevede

Thèse de doctorat sur travaux
de l'université paris-saclay

NNT : 2015SACLS007

préparée à Gustave Roussy

UMR1170 Hématopoïèse Normale et Pathologique

École Doctorale N° 582

Cancérologie : biologie - médecine - santé

Discipline : Aspects moléculaires et cellulaires de la biologie

Par

Jane Merlevède

Altérations génétiques et épigénétiques dans la leucémie myélomonocytaire chronique Modulation par les agents déméthylants

Thèse présentée et soutenue à Villejuif, le 1er octobre 2015

Directeur de thèse :	Éric Solary	Professeur des Universités - Praticien Hospitalier (Paris Sud)
Co-directeur de thèse :	Serge Koscielny	Biostatisticien (Gustave Roussy)
Composition du jury :		
Président du jury :	Daniel Gautheret	Professeur des Universités (Paris Sud)
Rapporteurs :	Daniel Birnbaum	Directeur de Recherche (INSERM, Marseille)
	Bruno Quesnel	Professeur des Universités - Praticien Hospitalier (Lille)
Examinateur :	Éric Letouzé	Chargé de Recherche (INSERM)

Remerciements

Pendant ces quatre années passées à Gustave Roussy, j'ai eu l'occasion de rencontrer de nombreuses personnes et je souhaite profiter de cette occasion pour remercier celles ayant contribué au déroulement de cette thèse par leur travail, leurs conseils ou leurs encouragements.

Je voudrais tout d'abord remercier mes deux directeurs de thèse Éric Solary et Serge Koscielny pour leur encadrement et le temps qu'ils m'ont consacré. Merci à Éric de m'avoir accueillie dans l'équipe pour ce doctorat même si mon parcours différait des us et coutumes du laboratoire. Merci à Serge d'avoir accepté cette co-direction même si mon sujet était différent de ses sujets de prédilections.

Un grand merci aux membres du jury pour avoir accepté d'évaluer cette thèse. Je remercie le Professeur Daniel Birnbaum et le Professeur Bruno Quesnel d'avoir accepté d'être rapporteurs, malgré des délais serrés. Merci également au Professeur Daniel Gautheret et au Docteur Éric Letouzé d'avoir accepté d'examiner ce travail.

Mes plus profonds remerciements vont à mon équipe pour leur contribution à ce travail. A Nathalie Droin tout d'abord pour l'important travail de séquençage d'échantillons d'exomes et de validations expérimentales de nombreux résultats ! Merci ensuite à Margot Morabito pour la préparation des échantillons analysés par séquençage. Merci aussi pour la collection des échantillons, c'est un travail de longue haleine. Je garderai le souvenir de nos multiples recherches d'échantillons aux -80° ! Un remerciement spécial à Stéphanie Solier pour sa précieuse aide tout au long de cette thèse, ses réponses à mes nombreuses questions, la relecture de cette thèse et son soutien dans les bons comme dans les mauvais moments. Je garderai en tête quelques moments mémorables ! Merci à Élisabeth Met et Nolwenn Lucas pour leur travail indispensable de récolte d'informations sur les patients. Merci également à Sagana et Séverine pour leur contribution. Merci à Julie Rivière et Laura Bencheikh pour leurs explications. Merci aux Masters qui sont passés par le bureau et qui ont eu droit également à pas mal de questions : Sarah, Aurélie et Camille.

Un très grand merci aux membres de l'unité. Aux chercheurs : William Vainchenker, Isabelle Plo, Fawzia Louache, Françoise Porteu, Jean-Luc Villeval,... pour leur aide et leurs conseils. Je garde en mémoire nombre de discussions et conseils. Aux jeunes chercheurs pour leur aide et sympathie : Sagana, Emna, Aurélie, Sébastien, Xénia, Mallorie, Marc, Mira, ... Merci à Elena Mylonas pour sa contribution à la première expérience de séquençage RNASeq. Bon courage à celles qui s'apprêtent à soutenir leur thèse.

Pendant ma thèse, j'ai pu compter sur l'expertise des membres de la plateforme de Bioinformatique. Merci à Marc Deloger et Khadija Diop pour leur aide précieuse dans les analyses de données et leur retour sur cette thèse. Un merci particulier pour Yannis Duffourd qui m'a épaulée aux débuts de cette thèse. Merci à Guillaume Meurice et Philippe Dessen pour leurs conseils. Merci à Pierre Bergoldt pour le support système. Merci à Céline Lefèbvre pour les relectures d'article et de thèse. Merci à Cathy Philippe pour ses conseils et rappels sur les nombreuses formalités de soutenance de thèse ! Merci aussi aux membres de la plateforme de Génomique, en particulier à Noémie Pata Merci et Marie Breckler, qui ont réalisé avec Nathalie la majorité des séquençages des données utilisées dans ce travail.

Je tiens à remercier nos collaborateurs sans qui tout ce travail n'aurait pas été possible. D'abord Thérèse Commes, qui nous a permis d'accéder aux données de séquençage de génomes de patients LMMC grâce à un financement France Génomique. Merci de m'avoir accueillie une semaine à Montpellier. Merci à son équipe : Florence Rufflé, Thomas Guignard, Anthony Boureux,... pour leur participation dans ce projet. Merci aux personnes du Centre National de Génotypage ayant réalisé le séquençage de ces échantillons et à Vincent Meyer pour son analyse de variations structurales. Puis, merci à Maria

Figueora et Tingting Qin pour le séquençage et l'analyse des données de méthylation, qui ont permis d'ajouter une importante plus-value à ce travail. Merci aussi à Émilie Chautard et Didier Auboeuf pour leur analyse des événements d'épissage. Enfin, merci à Seishi Ogawa et Kenichi Yoshida pour nous avoir fourni des données d'exomes.

Merci bien sûr aux patients qui acceptent d'être prélevés pour nous fournir du matériel d'étude et au Groupe Francophone des Myélodysplasies qui nous envoie les échantillons de patients. Je souhaite aussi remercier les personnes qui ont bien voulu me donner leur sang pour servir de contrôles : Philippe Dessen, Jean-Claude Ehrhart, Françoise Wendling et Alain Delain.

Mes pensées vont également aux membres de la plateforme Transcriptome et Épigénome de l'Institut Pasteur qui m'ont conseillée et aidée à mener à bien nos deux expériences RNA-Seq. Merci beaucoup à Hugo Varet pour son aide dans l'utilisation des contrastes avec DESeq2 ! Merci à Marie-Agnès Dillies pour ses conseils en analyse de données et en plan d'expérience. Merci de m'avoir initiée aux analyses de données NGS ! Merci également à Jean-Yves Coppée et Odile Sismero.

Merci aux organismes de financement qui m'ont permis d'aller au bout de cette thèse. Merci au comité de l'École Doctorale de Cancérologie de m'avoir accordé une chance en m'attribuant une bourse. Merci également à la Fondation pour la Recherche Médicale qui a financé 6 mois de ma dernière année (FDT20140931007).

Merci aux membres, anciens et actuels, de l'association de jeunes chercheurs de Gustave Roussy, avec qui j'ai passé de bons moments. J'espère que cette association va perdurer et attirer bon nombre de jeunes chercheurs.

Pendant les 3 premières années, j'ai mené une activité de médiation scientifique au Palais de la Découverte. Je garde un excellent souvenir de cette expérience où j'ai pu rencontrer des personnes de tous horizons. Je tiens à remercier les membres du département de Mathématiques pour cette expérience très enrichissante : Guillaume Reuiller, Pierre Audin, Romain Attal et Robin Jamet. J'ai également une pensée pour Aurélie Mabille, avec qui j'ai fait cette activité deux ans.

Je termine par un grand merci à mes proches. À mes amis pour leur soutien et leur encouragement pendant cette thèse : Naïla, Pierre, Jérôme, Marie, Michal, Arsène, aux "M2",... Un très grand merci à Pauline, qui me supporte depuis bientôt 20 ans ! Merci pour ton soutien durant cette folle aventure et merci pour la relecture de cette thèse ! On va enfin pouvoir profiter de ta vie parisienne ! Mille mercis à Antoine ! Merci de m'avoir supportée durant cette épreuve et surtout cette dernière année. C'est promis, fini les graphes en TikZ passé minuit ! Enfin, merci à ma famille pour votre soutien. Merci à Barbara, Johan, Anaïs ainsi qu'à mes neveux et nièces : Stanislas, Maxime, Fany, Louis, Sarah, Corentin, Victor et Mélie. Merci à tous ceux qui m'ont soutenue par le passé. Merci à ma mère de m'avoir indirectement conduite jusqu'ici et de m'avoir inculqué des valeurs qui me sont chères. Merci de m'avoir soutenue et encouragée à mieux faire.

Pour conclure, merci à toutes les personnes qui ont fait de ces années une expérience professionnelle enrichissante et merci à celles et ceux qui m'ont sincèrement soutenue dans les moments difficiles comme dans les bons. Merci à tous ceux qui m'ont encouragée et aidée à préparer la fin de cette thèse ces dernières semaines.

Je dédie ce travail aux patients atteints de LMMC.

Le Kremlin-Bicêtre, le 18 Août 2015.

Choisissez un travail que vous aimez et vous n'aurez pas à travailler un seul jour de votre vie.

Confucius

La théorie, c'est quand on sait tout et que rien ne fonctionne. La pratique, c'est quand tout fonctionne et que personne ne sait pourquoi. Ici, nous avons réuni théorie et pratique : Rien ne fonctionne... et personne ne sait pourquoi !

Albert Einstein

Table des matières

Liste des figures

Liste des tableaux

Glossaire

$\Delta\psi$ Percent Spliced In.
ACP Analyse en Composantes Principales.
ADNc ADN complémentaire.
ADNg ADN génomique.
ADN Acide DésoxyriboNucléique.
ARNm *ARN* messager.
ARN Acide RiboNucléique.
CLP Commun Lymphoid Progenitor.
CMP Common Myeloid Progenitor.
CNV Copy Number Variation.
CSH Cellule Souche Hématopoïétique.
DNMTi DNA MethylTransferase inhibitor.
ENCODE Encyclopedia of DNA Elements.
ERRBS Extended Reduced Representation Bisulfite Sequencing.
ESP NHLBI GO Exome Sequencing Project.
FDR False Discovery Rate.
FET Fisher Exact Test.
GLM Generalized Linear Model.
GM-CSF Granulocyte-Macrophage Colony-Stimulating Factor.
GMP Granulocyte Macrophage Progenitor.
INDEL Insertion - Délétion.
INSEE Institut National de la Statistique et des Études Économiques.
LMMC Leucémie MyéloMonocytaire Chronique.
LOH Loss of heterozygoty.
MDSC Myeloid Derived Suppressor Cells.
MPP MultiPotent Progenitor.
NGS Next Generation Sequencing.
OMS Organisation Mondiale de la Santé.
PCR Polymerase Chain Reaction.
RNASeq RNA Sequencing.
RRBS Reduced Representation Bisulfite Sequencing.
SNP Single Nucleotide Polymorphism.
SNV Single Nucleotide Variation.
WES Whole Exome Sequencing.
WGS Whole Genome Sequencing.
dbSNP Single Nucleotide Polymorphism database.

Préambule

La leucémie myélomonocytaire chronique est une maladie clonale de la cellule souche hématopoïétique qui touche le plus souvent des personnes âgées, l'âge moyen au diagnostic étant de 72 ans. C'est une pathologie rare puisque sa prévalence adaptée à l'âge est d'environ 0,4 individu sur 100000. La survie médiane est de moins de 3 ans après le diagnostic. Le décès survient du fait des conséquences de cytopénies (hémorragies ou infections), parfois dans le contexte d'une transformation en leucémie aiguë myéloïde. La seule thérapie curative est la greffe de cellules souches hématopoïétiques allogéniques, mais celle-ci n'est que rarement possible en raison de l'âge des patients. Les traitements actuellement utilisés sont l'hydroxyurée et les agents déméthylants. Ils retardent l'évolution de 40% des formes les plus sévères, sans jamais permettre la guérison.

Afin de comprendre les mécanismes en jeu dans cette pathologie, plusieurs pistes ont été récemment explorées. L'hypersensibilité des progéniteurs hématopoïétiques aux cytokines est une caractéristique générale des néoplasmes myéloprolifératifs. L'hypersensibilité de ces progéniteurs au Granulocyte-Macrophage Colony-Stimulating Factor (GM-CSF) a été décrite comme une des caractéristiques majeures de la leucémie myélomonocytaire juvénile, une pathologie pédiatrique, appartenant au même groupe d'hémopathies que la leucémie myélomonocytaire chronique et dans laquelle, le plus souvent, une mutation somatique provoque une activation constitutive de la voie RAS. Dans la leucémie myélomonocytaire chronique, l'hypersensibilité au *GM-CSF* est un peu moins systématique que dans la leucémie myélomonocytaire juvénile, mais elle est présente dans 40 à 80% des cas selon les critères utilisés pour la définir. Une autre caractéristique de la leucémie myélomonocytaire chronique est une dominance clonale précoce. Dans certains cas au moins, celle-ci pourrait participer au phénotype généré puisque la diminution de l'expression du gène *TET2*, le gène le plus souvent muté dans cette maladie, dans les cellules $CD34^{+}$, $CD38^{-}$ provoque un biais de différentiation vers la lignée granulo-monocytaire. Plusieurs autres altérations génétiques (la duplication interne en tandem dans *FLT3*, FLT3-ITD par exemple) ou épigénétiques (extinction de l'expression du gène *TIF1*γ par méthylation de son promoteur par exemple) peuvent induire une pathologie semblable à la leucémie myélomonocytaire chronique chez les souris vieillissantes.

Les études cytogénétiques identifient des anomalies acquises, de nombre et de structure, des chromosomes dans les cellules leucémiques de 30 à 50% des patients. Les pertes d'hétérozygotie sont très fréquentes, identifiées chez 50% des patients environ. Des mutations somatiques, perte ou gain de fonction, ont été identifiées dans une trentaine de gènes candidats ayant fait l'objet de cribles. Ces mutations affectent quatre familles de gènes codants des régulateurs épigénétiques, notamment *TET2* et *ASXL1* mutés chez $\simeq$ 60% et $\simeq$ 40% des patients respectivement, mais aussi *DNMT3A* (10-15%), *IDH2* et *EZH2* (<5%). Elles affectent aussi des composants des complexes appelés cohésines, comme *STAG2* (Kon et al. (2013)). Les facteurs d'épissage sont également fréquemment mutés avec *SRSF2* altéré chez un patient sur deux environ, *U2AF1*, *SF3B1* et *ZRSF2* chez moins de 15% des patients. Des protéines de la signalisation intra-cellulaire, en particulier de la voie RAS (*CBL*, *NRAS*, *KRAS*, *SH2B3*, *RIT1*), mais aussi d'autres voies (*JAK2*, *FLT3*) sont souvent mutées. Ces anomalies sont décrites en détails dans le chapitre 2. Des facteurs de transcription, principalement *RUNX1* (15% des patients), sont également altérés. L'analyse d'Itzykson et al. (2012) de 18 de ces gènes mutés dans la leucémie myélomonocytaire chronique conduite sur 312

patients a mis en évidence au moins une mutation somatique chez plus de 95% des patients.

Dans le cadre de ce doctorat, nous avons choisi d'étendre l'étude des anomalies génomiques et épigénomiques dans la leucémie myélomonocytaire chronique en analysant l'ensemble des séquences codantes du génome des cellules malades chez 49 patients et l'ensemble du génome des cellules malades chez 17 patients. Les cellules leucémiques étudiées ont été majoritairement des monocytes triés, ou dans quelques cas des cellules de la moelle osseuse ou des cellules mononuclées du sang. Des lymphocytes T, des fibroblastes cutanés ou des cellules de frottis de la muqueuse buccale ont été utilisés comme contrôles. Nous avons ensuite réalisé une étude longitudinale chez 17 patients traités ou non par des agents déméthylants et nous avons analysé l'effet de ces traitements sur les marqueurs épigénétiques chez 9 de ces patients. Enfin, nous avons analysé l'acide ribonucléique (ARN) de 10 patients et 4 sujets sains afin de déterminer les gènes différentiellement exprimés et différentiellement épissés dans les monocytes des patients.

Ces travaux ont été possibles grâce au développement du séquençage à très haut débit au cours de la dernière décennie. Cette technologie permet d'identifier d'éventuelles altérations de l'acide désoxyribonucléique (ADN) et de l'*ARN* dans les cellules malades comparées à des cellules saines. L'analyse de l'*ADN* permet la détection de mutations somatiques ou germinales à l'origine de diverses pathologies. L'analyse de l'*ARN* permet notamment de mesurer le niveau d'expression des gènes, d'identifier les sites de démarrage de la transcription et les divers épissages des ARN pré-messagers ou encore de détecter certaines anomalies de structure. Ces approches peuvent désormais être réalisées à l'échelle unicellulaire et dans des "biopsies liquides" à partir des acides nucléiques collectés dans un liquide biologique. Les méthodes d'analyse des séquences générées sont en plein développement.

Nous avons appliqué ces technologies à l'étude des altérations de l'*ADN* et de l'*ARN* dans les monocytes de patients atteints de leucémie myélomonocytaire chronique. Mon principal rôle a été de mettre en place et d'utiliser les outils nécessaires à l'analyse bio-informatique de ces données. Nous avons identifié 14 mutations géniques somatiques dans les régions codantes du génome et 475 dans l'ensemble des régions non répétées du génome en moyenne par patient. Nous avons établi le paysage mutationnel de la leucémie myélomonocytaire chronique et analysé l'impact des agents déméthylants dans ce contexte. Nos résultats suggèrent un effet plus épigénétique que cytotoxique des agents déméthylants, médicaments qui n'ont, au mieux, qu'un effet transitoire d'amélioration de l'hématopoïèse. Ces résultats décrits chapitre 5 ont été soumis à publication et sont en cours de révision. Nous avons également identifié plusieurs centaines de dérégulations géniques et anomalies d'épissage. Ces résultats sont présentés chapitre 6. L'ensemble de ces résultats est introduit par un état des lieux de la maladie étudiée, une introduction à la technologie utilisée et la description des analyses réalisées. Ces résultats sont ensuite discutés et mis en perspective.

Première partie

Introduction générale

Le cancer est la première cause de mortalité en France, juste devant les maladies cardio-vasculaires. Alors que le taux de décès par maladies vasculaires a diminué de moitié ces 25 dernières années, le taux de décès par cancer n'a que peu diminué et son incidence ne fait qu'augmenter. Chaque année, environ 350000 nouveaux cas sont détectés, soit 1 personne sur 220. On dénombre 150000 décès par an environ, soit 1 personne sur 430 environ. En 2012, on comptait environ 14 millions de nouveaux cas et 8,2 millions de décès dans le monde. Le nombre de nouveaux cas devrait augmenter de 70% environ au cours des deux prochaines décennies selon l'Institut National de la Statistique et des Études Économiques (INSEE). Le coût est non seulement humain mais aussi économique, de l'ordre de plusieurs milliards d'euros par an pour les soins et la recherche.

Du fait de notre patrimoine génétique, nous ne présentons pas tous les mêmes risques face aux cancers : il existe une composante génétique héréditaire dans au moins 10% des cancers. Les hommes et femmes ne sont pas affectés de la même manière. De manière générale, les hommes développent davantage de cancers, même si l'écart a diminué ces 20 dernières années. Le ratio homme/femme varie de manière importante suivant les cancers. Des inégalités attribuables à l'environnement et aux habitudes alimentaires ont été observées : les cancers les plus répandus dépendent, entre autres, de la position géographique et des habitudes alimentaires. Les mélanomes par exemple, sont particulièrement fréquents dans les pays à fort taux d'ensoleillement, mais ils sont rares au Japon où il est coutume de s'en protéger. Les agents infectieux sont impliqués dans 15% des cancers environ : le papillomavirus dans le cancer du col de l'utérus, le virus d'Epstein-Barr dans le lymphome de Burkitt ou la bactérie *Helicobacter pylori* dans le cancer gastrique. L'émergence d'une tumeur peut aussi être favorisée par des facteurs toxiques environnementaux : environ 20% de tous les cancers selon l'Organisation Mondiale de la Santé (OMS). Les inégalités économiques et sociales jouent également un rôle. Les pays en voie de développement ont le taux le plus élevé de décès par cancer. Aujourd'hui, plus de 60% des nouveaux cas de cancer et plus de 70% des décès par cancer surviennent en Afrique, Asie, Amérique centrale et Amérique latine selon l'*INSEE*. Les personnes à faible revenu semblent présenter un taux de mortalité plus élevé, même en France.

Selon l'*INSEE*, 30% des décès par cancer sont dus à cinq facteurs de risque : un indice élevé de masse corporelle, une faible consommation de fruits et légumes, le manque d'exercice physique, le tabagisme et la consommation d'alcool. Mais le hasard semble jouer également un rôle important puisque l'incidence des cancers dans un tissu donné est corrélé à la fréquence des divisions des cellules souches de ces tissus. Des dommages accidentels de l'*ADN* surviennent au cours de ces divisions. Les lésions non réparées peuvent contribuer à l'émergence de cancers.

Le cancer est une maladie caractérisée par une prolifération cellulaire anormale et/ou par un défaut de mort cellulaire au sein d'un tissu. Les cellules d'une tumeur sont toutes issues d'une même cellule anormale : elles forment donc un clone unique. En réalité, de nouvelles mutations s'accumulent au cours de l'évolution et définissent des sous-clones qui, dans certaines tumeurs en fin d'évolution, génèrent une hétérogénéité génétique intra-tumorale importante, source de résistance au traitement. La cellule d'origine du cancer doit avoir acquis, du fait d'une ou plusieurs mutations, un avantage compétitif vis-à-vis des cellules voisines.

La découverte de clones de cellules mutées dans le sang ou la peau de sujets sains suggère que la survenue d'une mutation puis l'émergence d'un clone ne suffisent pas à définir un cancer. Il faut que l'environnement, devenu permissif, favorise l'expansion des cellules du clone de façon incontrôlée : les cellules du clone se multiplient, deviennent insensibles aux signaux régulateurs notamment anti-prolifératifs, meurent moins facilement, génèrent la formation de vaisseaux et

parfois migrent à distance du tissu d'origine pour former des métastases.

Les cancers sont classés en deux groupes :
— les cancers "solides", comprenant les carcinomes et les sarcomes
— les cancers "liquides" ou sanguins, comprenant les leucémies et les lymphomes

Les carcinomes se développent à partir d'un épithélium, les sarcomes à partir de tissus conjonctifs, comme les os, et les cancers sanguins à partir des cellules souches ou des progéniteurs hématopoïétiques. Les cancers hématopoïétiques représentent moins de 10% des cancers. Ils peuvent être aigus ou chroniques. Les hémopathies myéloïdes chroniques sont classées en néoplasmes prolifératifs, syndromes myélodysplasiques et formes frontières.

Nous nous intéressons dans ce travail à la leucémie myélomonocytaire chronique qui est classée par l'*OMS* dans un groupe de formes frontières entre néoplasmes prolifératifs et syndromes myélodysplasiques. La leucémie myélomonocytaire chronique est la plus répandue de ces formes frontières.

Leucémie myélomonocytaire chronique 1

Dans ce chapitre, nous décrivons la leucémie myélomonocytaire chronique, ses critères diagnostic, son expression clinique et ses anomalies biologiques. Les mutations somatiques identifiées dans la leucémie myélomonocytaire chronique font l'objet d'un chapitre à part entière. La leucémie myélomonocytaire chronique étant une anomalie de l'hématopoïèse, nous décrivons cette dernière, lors de son fonctionnement normal et lors d'hémopathies. Ce chapitre s'achève sur les traitements actuels de la leucémie myélomonocytaire chronique et leur mode d'action.

1.1 Qu'est ce que la leucémie myélomonocytaire chronique selon l'Organisation Mondiale de la Santé ?

L'entité leucémie myélomonocytaire chronique a été identifiée par le groupe FAB (French/American/British) en 1994 à partir de la monocytose (Bennett et al. (1994)) et validée par l'*OMS* en 2000 (Bennett (2000)). La dernière mise à jour date de 2008 (Vardiman et al. (2009)), une autre est en préparation. Le groupe en question a été appelé "Néoplasmes myéloprolifératifs / Syndromes myélodyplasiques" en 2008. La leucémie myélomonocytaire chronique est l'entité la plus fréquente dans ce groupe de maladies.

Il s'agit d'une maladie clonale de la cellule souche hématopoïétique (CSH) ou d'un progéniteur très immature dont le critère diagnostique principal (et le seul critère positif officiel) est une

monocytose sanguine supérieure à 1G/L persistante pendant au moins 3 mois.

Selon l'*OMS*, le diagnostic de leucémie myélomonocytaire chronique impose de valider trois critères négatifs :

— Il ne doit pas y avoir de chromosome Philadelphie au caryotype ou de réarranegement BCR-ABL à la réaction en chaîne (PCR, Polymerase Chain Reaction) afin d'éliminer la leucémie myéloide chronique
— Le pourcentage de blastes dans la moelle et le sang doit être inférieur à 20% afin d'éliminer une leucémie aiguë
— Il ne doit pas exister d'hyperéosinophilie supérieure à $1500/mm^3$ qui est habituellement le reflet d'une translocation chromosomique équilibrée affectant notamment le gène *PDGFRα*. Ces hémopathies entrent désormais dans le cadre des néoplasmes myéloïdes et lymphoïdes avec éosinophilie et sont sensibles à l'Imatinib.

La présence d'une dysplasie cellulaire touchant au moins une lignée myéloïde est un argument diagnostique utile mais inconstant et parfois subjectif. Le myélogramme montre une moelle de richesse souvent normale ou augmentée. La présence de monocytes et de promonocytes est caractéristique. L'*OMS* distingue la leucémie myélomonocytaire chronique de type 1, avec moins de 5% de blastes dans le sang périphérique et moins de 10% de blastes dans la moelle osseuse, de la leucémie myélomonocytaire chronique de type 2 avec entre 5 et 19% de blastes dans le sang périphérique et de 10 à 19% de blastes médullaires.

Comment améliorer le diagnostic de leucémie myélomonocytaire chronique ?

La dysplasie n'est pas toujours facile à identifier. De ce fait, les cytologistes considèrent souvent que le diagnostic de leucémie myélomonocytaire chronique est délicat et que d'autres marqueurs seraient nécessaires.

La biopsie médullaire n'est pas souvent réalisée en France, elle l'est beaucoup plus dans les autres pays dans lesquels elle est un critère d'évaluation de la réponse au traitement. Outre la richesse médullaire et l'infiltration monocytaire, elle permet d'évaluer la fibrose et de noter la présence fréquente d'amas de cellules dendritiques plasmacytoïdes CD123 positives dont la signification est indéterminée.

Le caryotype médullaire peut apporter un argument complémentaire solide : la présence d'une anomalie cytogénétique acquise et clonale distingue la leucémie myélomonocytaire chronique de la monocytose réactionnelle. Une telle anomalie, qui n'est jamais spécifique (trisomie 8, perte du chromosome Y, anomalie du chromosome 7 sont les plus fréquentes), n'est observée que dans 30 à 40% des cas (Such et al. (2011)). Les résultats du caryotype sont indicatifs de la gravité de la maladie (Wassie et al. (2014)) et la détection de nouvelles anomalies cytogénétiques en cours d'évolution, observée chez 25% des patients, est un marqueur de l'aggravation de la maladie (Tang et al. (2015)).

La biologie moléculaire peut apporter un argument complémentaire solide : la présence d'une mutation somatique affectant un gène muté de façon récurrente dans les hémopathies myéloïdes (une centaine de gènes identifiés dans ces pathologies) est également en faveur d'une leucémie myélomonocytaire chronique par rapport à une monocytose réactionnelle.

La cytométrie en flux apporte une ou plusieurs informations supplémentaires. Notre équipe

a détecté, chez les patients atteints de leucémie myélomonocytaire chronique, une anomalie de la répartition des trois sous-populations de monocytes circulants : les monocytes classiques $CD14^+$,$CD16^-$ sont augmentés aux dépens des monocytes intermédiaires $CD14^+$,$CD16^+$ et des monocytes non classiques $CD14^-CD16^+$. Lorsque la fraction des monocytes classiques $CD14^+$,$CD16^-$ représente plus de 94% des monocytes circulants, le diagnostic de leucémie myélomonocytaire chronique devient hautement probable, ce test étant à la fois très sensible et très spécifique (Selimoglu-Buet et al. (2015)). Des travaux en cours au sein de l'équipe cherchent à identifier l'origine de cette anomalie de répartition que les agents déméthylants semblent corriger lorsqu'ils induisent une réponse thérapeutique.

1.2 Quels sont les signes cliniques et biologiques de la maladie ?

Si la leucémie myélomonocytaire chronique est le plus fréquent des néoplasmes myéloprolifératifs / syndromes myélodysplasiques, elle reste une maladie rare qui représente moins de 2% des hémopathies malignes de l'adulte. Son incidence adaptée à l'âge, récemment révisée, est de 0,4 / 100000 / an. Son incidence augmente avec l'âge. L'âge médian au diagnostic est compris entre 70 et 75 ans. Le diagnostic est exceptionnel avant 50 ans. Le pronostic semble meilleur lorsque l'âge au diagnostic est inférieur à 65 ans (Patnaik et al. (2015)). La maladie touche 2 fois plus souvent les hommes que les femmes.

Le diagnostic de leucémie myélomonocytaire chronique est souvent fortuit car la maladie est asymptomatique ou induit des signes cliniques peu spécifiques à ses débuts : il s'agit le plus souvent de la conséquence de cytopénies (fatigue, infection ou hémorragie) et de signes généraux comme la fièvre, les sueurs nocturnes ou l'amaigrissement. Elle est susceptible d'induire une splénomégalie ou plus rarement une hépatomégalie. Les signes cliniques extra-hématopoiétiques sont rares, à l'exception des lésions infiltratives cutanées. Des localisations neurologiques ou digestives ont été rapportées. Les manifestations auto-immunes associées sont très classiques et affectent 10 à 15% des patients.

L'hémogramme révèle la monocytose, par définition au moins égale à 1G/L, parfois bien plus élevée. Il existe fréquemment une polynucléose, parfois une myélémie et des cellules granuleuses immatures ou dysplasiques douées de fonctions immunosuppressives. Il peut exister une thrombocytose modérée au diagnostic. Plus souvent, il existe une thrombopénie et une anémie de degrés variables, parfois importantes et nécessitant des transfusions répétées. L'analyse des lymphocytes circulants montre souvent une augmentation des lymphocytes T régulateurs.

Du fait de la monocytose, le lyzozyme sérique et urinaire est augmenté. Du fait de la composante proliférative, le taux plasmatique des lactates déshydrogénases, de la vitamine B12, de l'uricémie peut être augmenté. Ces examens sont de peu d'utilité pratique.

L'imagerie a souvent peu d'utilité. L'échographie abdominale précise la taille de la rate et du foie. La tomoscintigraphie par émission de positons peut être utilisée pour identifier un foyer d'hématopoïèse extra-médullaire, une éventualité très rare.

La médiane de survie est de deux à trois ans après le diagnostic. Le décès survient du fait de complications infectieuses ou hémorragiques, parfois dans le contexte d'une transformation en leucémie aiguë myéloïde.

1.3 Quelle est notre compréhension actuelle de la physiopathologie de la leucémie myélomonocytaire chronique ?

La leucémie myélomonocytaire chronique est décrite comme une maladie clonale de la cellule souche hématopoïétique. Avant d'explorer les hypothèses physiopathologiques, nous faisons un résumé, nécessairement succinct, de notre compréhension actuelle de l'hématopoïèse normale.

1.3.1 Les grands principes de l'hématopoïèse normale

L'hématopoïèse est l'ensemble des mécanismes qui assurent le remplacement continu et régulé de l'ensemble des cellules sanguines. Chez l'Homme, et plus généralement chez les vertébrés, elle connaît deux phases au cours du développement. La première, dite transitoire, constitue l'hématopoïèse primitive. Elle se déroule à l'extérieur de l'embryon, dans le sac vittelin composé de cellules dérivées du mésoderme et de l'endoderme. Sa fonction principale est de produire rapidement des globules rouges pour oxygéner les tissus, des macrophages pour modeler les tissus et des mégacaryocytes pour permettre le développement des vaisseaux de l'embryon. La seconde constitue l'hématopoïèse définitive. Elle se met en place au bout de 27 à 40 jours de développement à l'intérieur de l'embryon, migrant de territoire en territoire : de l'AGM (Aorte, Gonades, Mésonéphros) au foie fœtal, au thymus et à la rate pour s'établir définitivement dans la moelle osseuse chez l'Homme adulte.

Le système hématopoïétique est organisé hiérarchiquement. Au sommet de la hiérarchie se trouvent les *CSH* multipotentes. Elles ont une durée de vie longue, plusieurs années et en théorie infinie, et sont localisées dans les niches de la moelle osseuse. Elles sont capables aussi bien d'autorenouvellement, *i.e.* de donner une cellule fille équivalente, que de différenciation, *i.e.* elles sont à l'origine de toutes les cellules sanguines myéloïdes et lymphoïdes. L'autorenouvellement préserve ce compartiment de *CSH* multipotentes rares (moins de 1 cellule sur 10000 dans la moelle osseuse) et quiescentes, *i.e.* ces cellules se divisent très rarement au cours de la vie. Ces cellules ont également des capacités de migration : elles migrent dans le sang afin de coloniser les organes hématopoïétiques. L'équilibre entre autorenouvellement et engagement vers une voie de différenciation spécifique est capital.

Cette organisation complexe aurait pu disparaître à l'issue de la phase de développement de l'organisme mais elle s'est maintenue à l'âge adulte. Certains auteurs considèrent que ce maintien est la conséquence d'un processus darwinien de sélection, un tel système diminuant considérablement le risque de cancer, puisque seules les *CSH* sont celles dont la transformation aura les conséquences les plus délétères (Pepper et al. (2007)).

Lorsque la *CSH* se différencie, elle génère un progéniteur. Celui-ci peut avoir une grande capacité de prolifération et de différenciation mais a perdu la capacité d'autorenouvellement. À partir de ce progéniteur, chaque étape de la différenciation s'accompagne de la génération de progéniteurs de plus en plus engagés dans la différentiation en lignées spécifiques. Un Progéniteur MultiPotent (MPP) donne un Progéniteur Commun Myéloïde (CMP), aussi appelé CFU-GEMM (Colony Forming Unit) et un Progéniteur Commun Lymphoïde (CLP) (figures 1.1 et 1.2). Le *CLP* donne toutes les lignées lymphoïdes. Ces lignées protègent l'organisme par une réponse immunitaire acquise, médiée par les lymphocytes B et T et par une réponse immunitaire innée, médiée notamment par les lymphocytes NK. Le *CMP* donne un progéniteur commun aux lignées mégacaryocytaire et érythroblastique MEP (ou E/Mk) et un progéniteur commun aux lignées granuleuse et monocytaire (Granulocyte Macrophage Progenitor (GMP) ou CFU-GM) (figure 1.2).

Les précurseurs sont des cellules qui ont commencé la synthèse de protéines spécifiques

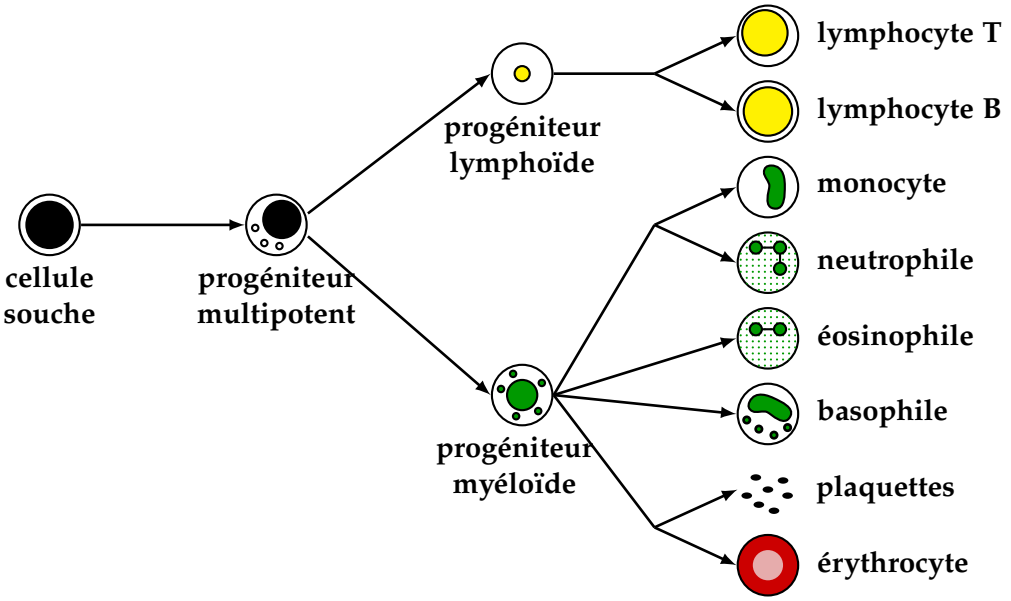


FIGURE 1.1 – *Schéma de l'hématopoïèse (adapté de http ://wwwasagadospacientesmm.blogspot.fr/2012/07/v-behaviorurldefaultvmlo.html)*

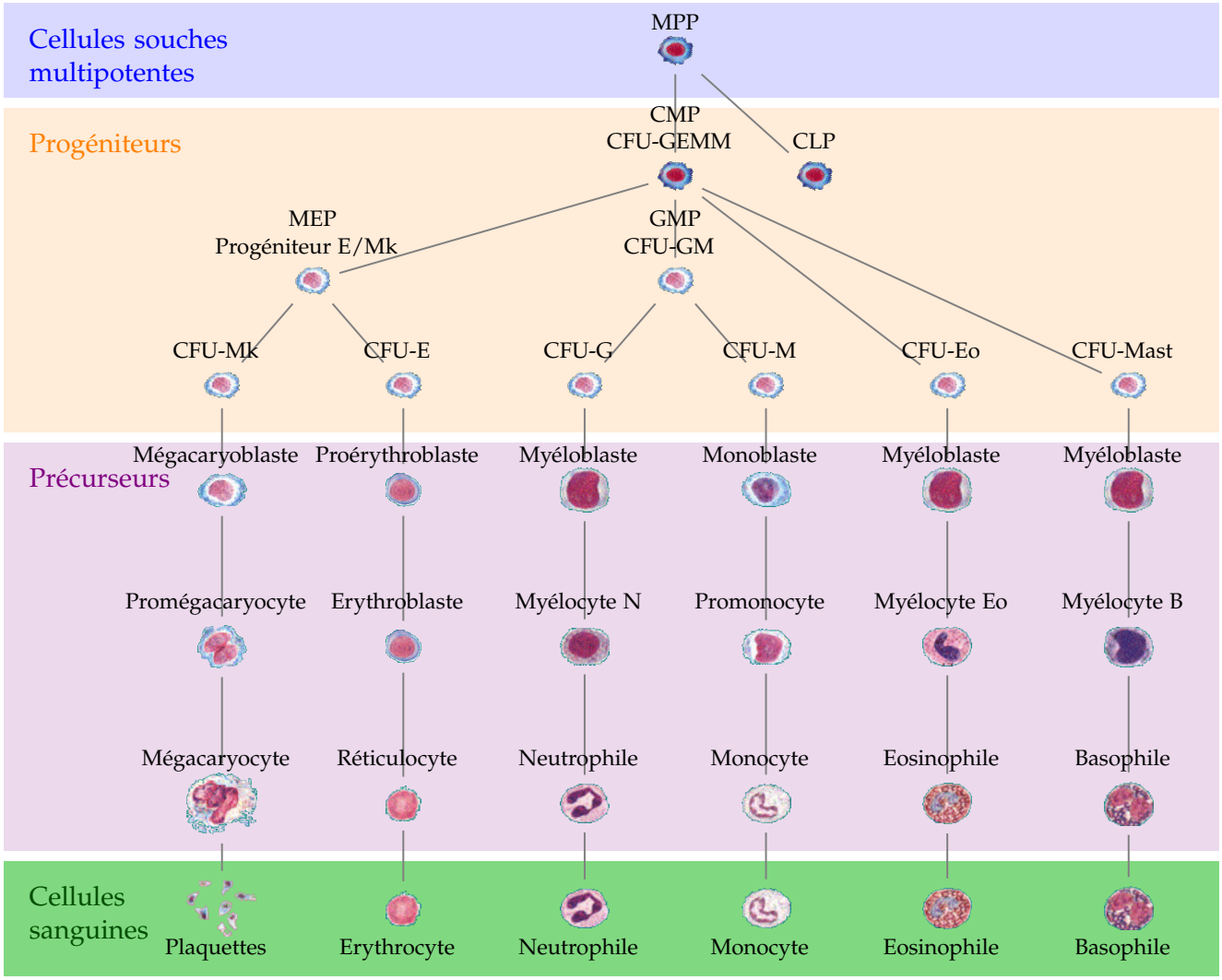


FIGURE 1.2 – *Principaux acteurs de l'hématopoïèse myéloïde (adapté de http ://www.toutsurlatransfusion.com/transfusion-sanguine/medecine-transfusionnelle/notions-d-hematologie.php)*

de lignée et sont identifiables morphologiquement. Ils ont des capacités de prolifération très limitées : trois mitoses maximum. Les précurseurs les plus matures sont les mégacaryocytes, les réticulocytes, les neutrophiles, les monocytes, les éosinophiles et les basophiles (figure 1.2). Les monocytes conservent une capacité de différenciation lorsqu'ils migrent dans les tissus.

La dernière étape de l'hématopoïèse consiste en la sortie des précurseurs les plus matures de la moelle osseuse vers le sang. Les cellules matures de la lignée myéloïde présentes dans le sang sont les érythrocytes, plaquettes, monocytes, polynucléaires basophiles, éosinophiles et neutrophiles. Les plaquettes sont produites directement dans le sang par fragmentation du cytoplasme de leur précurseur et permettent la coagulation. Les érythrocytes assurent le transport de l'oxygène. Les macrophages participent à l'immunité innée et parfois acquise en phagocytant les débris cellulaires et les pathogènes. Les polynucléaires sont impliqués dans la réponse aux infections. La plupart des cellules matures du sang circulant ont une durée de vie courte : quelques heures pour les polynucléaires neutrophiles, quelques jours pour les monocytes et polynucléaires basophiles, 10 jours pour les plaquettes, 120 jours pour les érythrocytes. Ces cellules ne se multiplient pas, donc le processus d'hématopoïèse doit assurer le renouvellement continu de grandes quantités de cellules sanguines matures. Chez l'Homme, plus de 2.10^{11} cellules sanguines sont produites chaque jour. Ce système de production est régulé de telle sorte que le nombre de cellules du sang varie peu dans les conditions normales. Ce système est capable de répondre à de nombreux stress comme une infection ou une hémorragie en modulant la production de cellules spécifiques.

Les détails de ce schéma "classique" de l'hématopoïèse sont appelés à être révisés. Les techniques de barcoding (Perié et al. (2014)) traçant l'évolution des cellules à l'échelle unicellulaire permettent désormais de ré-analyser ce schéma et d'identifier de nouvelles voies de maturation des cellules sanguines au sein de la moelle osseuse.

1.3.2 Les grandes caractéristiques de l'hématopoïèse pathologique

Les anomalies de l'hématopoïèse consistent en une dérégulation du nombre (excès ou insuffisance) ou de la qualité (dysfonctionnement) des cellules produites.

On distingue les hémopathies bénignes des hémopathies malignes. Parmi les hémopathies bénignes, on peut citer les conséquences de carences (fer et certaines vitamines), les maladies infectieuses (en particulier parasitaires, la plus répandue étant le paludisme) ou les maladies génétiques comme l'hémophilie et les hémoglobinopathies. Les deux hémoglobinopathies les plus courantes en France sont la thalassémie et la drépanocytose. Cette dernière est la maladie génétique la plus fréquente en France, elle consiste en une altération de la forme des globules rouges. La thalassémie est due à une insuffisance ou une absence de production d'une sous-unité de l'hémoglobine.

Les hémopathies malignes peuvent toucher soit le tissu lymphoïde, soit le tissu myéloïde et présenter une forme aiguë ou chronique. Les principales hémopathies lymphoïdes sont la leucémie aiguë lymphoblastique, la leucémie lymphoïde chronique, le lymphome malin non Hodgkinien, la maladie de Hodgkin et le myélome multiple des os (maladie de Kahler). Puisqu'ils n'affectent pas la même branche que notre pathologie d'intérêt, nous ne les détaillerons pas. Les hémopathies myéloïdes représentent environ 34% des hémopathies malignes et présentent également un aspect aigu et un aspect chronique.

Forme aiguë La leucémie aiguë myéloïde (ou myéloblastique) peut survenir à tout âge (25% des cas sont diagnostiqués avant 25 ans). Elle peut être primitive (leucémie *de novo*) ou secondaire,

notamment à une hémopathie chronique. C'est ainsi que 25 à 35% des leucémies myélomonocytaires chroniques se transforment en leucémie aiguë myéloïde. La leucémie aiguë myéloïde représente environ 7% des hémopathies malignes. Elle se traduit par un blocage de maturation des cellules médullaires à un stade souvent très immature. Il y a multiplication incontrôlée de blastes qui envahissent la moelle osseuse. Celle-ci n'assure plus la production des cellules sanguines normales, ce qui peut conduire à une anémie, une neutropénie et une thrombocytopénie. Actuellement traitée par des cures de polychimiothérapie intensive, on observe une réponse complète dans 50 à 99% des cas en fonction des patients et de leur maladie. Les rechutes sont cependant fréquentes et le pronostic est hétérogène : la guérison est obtenue dans plus de 95% des leucémies aiguës promyélocytaires par la combinaison acide tout-trans rétinoïque et trioxyde d'arsenic, mais elle est très rare dans les leucémies aiguës myéloïdes avec caryotype complexe, en particulier chez les sujets âgés. Les risques de rechute sont moindres si une allogreffe de moelle peut être effectuée, mais au prix d'une mortalité liée à la greffe qui augmente avec l'âge.

Formes chroniques Les hémopathies myéloïdes chroniques sont caractérisées par un nombre anormal de progéniteurs. Elles sont classées en trois catégories principales : les néoplasmes myéloprolifératifs, les syndromes myélodysplasiques et les syndromes mixtes qui représentent 12%, 11.8% et 3.2% respectivement des hémopathies malignes.

Les néoplasmes myéloprolifératifs se caractérisent par l'expansion clonale d'une ou plusieurs lignées myéloïdes. Ils présentent une production anormalement élevée d'un progéniteur. Le mécanisme physiopathologique dominant est une hypersensibilité des progéniteurs myéloïdes à une ou plusieurs cytokines. Cette hypersensibilité est la conséquence d'altérations génétiques acquises affectant les voies de signalisation cellulaire.

- la leucémie myéloïde chronique, caractérisée par la translocation t(9 :22), résultant en la protéine de fusion bcr-abl. Les malades sont surtout des personnes âgées de 20 à 50 ans qui présentent une production anormalement élevée de globules blancs. Le risque de transformation en leucémie aiguë myéloïde est élevé, mais la greffe de cellules souches hématopoïétiques, puis l'interféron alpha et surtout les inhibiteurs de tyrosine kinase comme l'Imatinib ont transformé le pronostic de cette maladie autrefois constamment fatale.
- la polyglobulie de Vaquez est une maladie du sujet âgé, caractérisée par une production anormalement élevée de globules rouges. Plus de 90% des malades possèdent une mutation dans le régulateur de signalisation *JAK2* (Janus kinase 2) (James et al. (2005)). Le risque d'évolution en leucémie aiguë myéloïde est compris entre 5 et 10% et en myélofibrose est d'environ 20% après 10 ans d'évolution (Czader and Orazi (2015)).
- la thrombocytémie essentielle peut apparaître à tout âge, avec une médiane d'âge au diagnostic de 60-65 ans. Les malades produisent trop de plaquettes. En plus de la mutation dans *JAK2* dans 50% des cas et des mutations de *MPL* dans 15% des cas, des mutations dans *CALR* (Calréticuline) ont été mises en évidence récemment (Nangalia et al. (2013)) chez 25-30% des patients. Le risque d'évolution en leucémie aiguë myéloïde est compris entre 2 et 5% (Czader and Orazi (2015)).
- la myélofibrose primitive est une maladie du sujet âgé. Environ 50% des malades possèdent une mutation dans *JAK2* et 20-30% dans *CALR*. La myélofibrose secondaire fait suite à une maladie de Vaquez ou à une thrombocytémie essentielle. Le risque d'évolution en leucémie aiguë myéloïde est d'environ 10% (Czader and Orazi (2015)).
- la mastocytose systémique est caractérisée par des mutations de *KIT*, la leucémie chronique à neutrophiles par des mutations de *CSF3R* et les hémopathies chroniques à éosinophiles par des mutations de *PDGFRβ*.

Les syndromes myélodysplasiques affectent souvent des gens plus âgés. Les mécanismes phy-

siopathologiques de la dysplasie sont moins clairs que ceux de la myéloprolifération. Les formes les moins agressives se caractérisent par une apoptose excessive des précurseurs ou des progéniteurs myéloïdes dans la moelle. Les formes plus agressives se rapprochent des leucémies aiguës myéloïdes. La classification *OMS* de 1999 répartit les syndromes myélodysplasiques en sept catégories (table 1.1). Des anomalies de la synthèse protéique dues à des altérations des ribosomes semblent contribuer à l'érythroblastopénie. Ces syndromes sont en fait la conséquence de l'accumulation d'un nombre élevé d'altérations géniques conduisant à diverses cytopénies plus ou moins sévères précédant la transformation en leucémie aiguë myéloïde.

	Sang	Moelle
Anémie réfractaire	Anémie Absence ou rares myéloblastes	Dysplasie érythroblastique uniquement <5% myéloblastes, <15% sidéroblastes en couronne
Cytopénie réfractaire avec dysplasie multilignée	Cytopénie Absence ou rares myéloblastes Monocytes<1G/L	Dysplasie $\geq$10% des cellules et $\geq$ 2 lignées myéloïdes <5% myéloblastes, <15% sidéroblastes en couronne Pas de corps d'Auer
Anémie réfractaire avec sidéroblastes en couronne	Anémie Absence de myéloblaste	Dysplasie érythroblastique isolée <5% myéloblastes, $\geq$15% sidéroblastes en couronne
Cytopénie réfractaire avec dysplasie multilignée et sidéroblastes en couronne	Bi ou Pan-cytopénie Absence ou rares myéloblastes Pas de corps d'Auer Monocytes<1G/L	Dysplasie $\geq$10% des cellules et $\geq$ 2 lignées myéloïdes <5% myéloblastes, $\geq$15% sidéroblastes en couronne Pas de corps d'Auer
Anémie réfractaire avec excès de blastes 1 AREB1	Cytopénies <5% myéloblastes Pas de corps d'Auer Monocytes<1G/L	Dysplasie uni ou multi-lignée 5 à 9% myéloblastes Pas de corps d'Auer
Anémie réfractaire avec excès de blastes 2 AREB2	Cytopénies <5 à 19% myéloblastes $\pm$ de corps d'Auer Monocytes<1G/L	Dysplasie uni ou multi-lignée 10 à 19% myéloblastes $\pm$ de corps d'Auer
SMD non classé	Cytopénies Absence ou rares myéloblastes Pas de corps d'Auer	Dysplasie unilignée (lignée granuleuse ou mégacaryocytaire) Pas de corps d'Auer <5% myéloblastes
SMD avec délétion 5q- isolée	Anémie <5% myéloblastes Plaquettes normales ou augmentées	Mégacaryocytes normaux ou augmentés avec noyaux hypolobés <5% myéloblastes Pas de corps d'Auer Délétion isolée (5q-)

TABLE 1.1 – *Classification OMS des syndromes myélodysplasiques*

Les néoplasmes myéloprolifératifs / syndromes myélodysplasiques ont été créés pour regrouper les pathologies mixtes. La leucémie myélomonocytaire en est l'exemple le plus fréquent, mais il en existe d'autres. Toutes ont en commun un pourcentage de blastes inférieur à 20% dans le sang et la moelle, l'absence de chromosome Philadelphie ou de réarrangement BCR-ABL, l'absence d'hyperéosinophilie supérieure à 1500/mm^3 et de réarrangement du gène *PDGFR*β.

— La leucémie myéloide chronique atypique est diagnostiquée chez les personnes âgées de plus de 65 ans en moyenne avec hyperleucocytose majeure et myélémie constante. Elle est souvent associée à des mutations de *SETBP1* et de *CSF3R* (Maxson et al. (2013)).

— La leucémie myélomonocytaire chronique juvénile survient chez les enfants de moins de 15 ans, le plus souvent des garçons de moins de 3 ans. Les signes cliniques sont une fièvre, une spléno-hépatomégalie (infiltration leucémique), une hyperleucocytose sanguine avec monocytose et petite myélémie, et des signes d'infection bactérienne et / ou virale. Dans plus de 90% des cas est retrouvée une mutation de la voie RAS (*NRAS*, *KRAS*, *NF1*, *PTPN11* ou

CBL). Une partie des leucémies myélomonocytaires juvéniles évolue favorablement après chimiothérapie peu agressive ou de façon spontanée de manière encore inexpliquée.
— Les néoplasmes myéloprolifératifs / syndromes myélodysplasiques dits inclassifiables regroupent les autres pathologies mixtes. Elles sont rares et encore mal identifiées.
— Les anémies réfractaires sidéroblastiques avec thombocytose associent une anémie sidéroblastique (par mutation de *SF3B1*) et une thrombocytose (par mutation de *JAK2*, *MPL* ou *CALR*).

1.4 Que savons-nous de la physiopathologie de la leucémie myélomonocytaire chronique ?

La cellule d'origine est un progéniteur très immature. Les anomalies moléculaires ont été détectées dans les cellules médullaires $CD34^+$,$CD38^-$,$CD90^+$. La majorité de ces cellules est mutée, indiquant que les cellules mutées envahissent le compartiment immature. Pour autant, toutes ces cellules n'ont pas le même statut mutationnel : on peut reconstituer l'acquisition des altérations génétiques en examinant le compartiment immature. Cela suggère que, dans ce compartiment, les altérations génétiques qui s'accumulent donnent sans doute à ces cellules un petit avantage par rapport aux cellules normales et que leur accumulation n'accroît pas cet avantage. En revanche, au cours de la différenciation, on observe très vite que seules les cellules les plus mutées sont amplifiées (Itzykson et al. (2013b)).

Dans le clone, une partie importante du compartiment lymphoïde T ne porte pas les mutations somatiques détectées dans les cellules myéloïdes, suggérant que la cellule d'origine est un progéniteur déjà engagé ayant retrouvé des propriétés d'autorenouvellement ou que les mutations induisent un désavantage compétitif lors de la différenciation lymphoïde.

1.4.1 Biais de différenciation vers la lignée granulomonocytaire

A partir des progéniteurs mutés, la différenciation est orientée vers la production excessive de cellules granulomonocytaires, au détriment des cellules érythro-mégacaryocytaires (Itzykson et al. (2013b)). Deux facteurs semblent contribuer à ce déséquilibre de l'hématopoïèse : l'accumulation de certaines mutations, comme les mutations "perte de fonction" de *TET2* et les mutations affectant les voies de signalisation. Elles pourraient sensibiliser les progéniteurs granulomonocytaires au *GM-CSF*, comme dans les néoplasmes myéloprolifératifs.

1.4.2 Sensibilité au Granulocyte-Macrophage Colony-Stimulating Factor

Cette hypersensibilité des progéniteurs au *GM-CSF* est une caractéristique importante des leucémies myélomonocytaires juvéniles, initialement décrite fin des années 80 (Gualtieri et al. (1989)). Environ 90% des patients présentent cette hypersensibilité.

Chaque cytokine liée au récepteur du *GM-CSF* lie des chaînes α spécifiques et partage une chaîne β commune nécessaire à l'activation. Les chaînes α et β du récepteur du *GM-CSF* se combinent pour former un hétérodimère, permettant l'association avec JAK2. L'interaction et la phosphorylation par JAK2 du récepteur sont nécessaires pour initier la signalisation intracellulaire qui conduit à la transduction du signal et à l'activation de la transcription STAT5 et Ras.

Puisque la signalisation du *GM-CSF* est essentielle pour la survie et la différentiation des monocytes, cibler le *GM-CSF* dans les traitements a été exploré avec plus ou moins de succès dans les leucémies myélomonocytaires juvéniles, leucémies aiguës myéloïdes et leucémies myélomonocytaires chroniques (Frankel et al. (1998)). Dans la leucémie myélomonocytaire chronique, environ

40% des patients présentent cette hypersensibilité au *GM-CSF*. Des traitements combinant le ciblage du *GM-CSF* et l'inhibition de JAK2 ont montré un potentiel thérapeutique (Padron et al. (2013)).

Si l'on s'intéresse ensuite à la partie circulante du clone, on retient deux anomalies. La première est un défaut d'apoptose des monocytes qui s'accumulent dans le sang des patients. Ce défaut d'apoptose semble être la conséquence d'une augmentation de l'expression de Mcl-1 et de Bcl-XL. L'augmentation de l'expression de ces protéines anti-apoptotiques de la famille Bcl-2 pourrait être une conséquence de l'hyperactivation de la voie STAT5. La seconde est la présence, en quantité variable, de cellules granuleuses immatures et dysplasiques douées de fonctions immunosuppressives, aussi bien vis-à-vis de l'immunité innée (blocage de la différentiation des monocytes en macrophages) que de l'immunité acquise (effet cytotoxique sur les lymphocytes T activés). Ces cellules se comportent comme les cellules myéloïdes suppressives (MDSC, myeloid derived suppressor cells) des immunologistes. La production d'un nombre élévé de ces cellules est un facteur de mauvais pronostic.

1.4.3 Présence de cellules granuleuses immatures et dysplasiques

Les *MDSC* ont été initialement décrites dans les années 70 (Strober (1984)) mais des limitations techniques ont retardé leur étude jusque dans les années 90. Depuis, elles sont largement étudiées. Les *MDSC* (Khaled et al. (2013), Trikha and Carson (2014), Jiang et al. (2014)) constituent un ensemble de cellules myéloïdes immatures, composé de progéniteurs, de macrophages, de granulocytes et de cellules dendritiques immatures qui s'accumulent dans le sang, les organes lymphoïdes, la rate et les tissus tumoraux dans plusieurs conditions pathologiques : infection, septicémie ou oncogénèse. La principale action des *MDSC* consiste en la régulation négative de la réponse immunitaire durant la progression tumorale, l'inflammation et l'infection, favorisant de ce fait la progression tumorale. Chez les sujets sains, les *MDSC* représentent 0.5% des cellules mononuclées du sang, alors qu'elles sont par exemple 10 fois plus présentes chez les patients atteints de cancers du rein ou du colon (Khaled et al. (2013)). Le nombre de *MDSC* dans le sang des patients est corrélé positivement à la charge tumorale et à l'état clinique (Jiang et al. (2014)). L'engouement pour les *MDSC* s'explique par l'éventualité d'améliorer l'index thérapeutique des chimiothérapies et thérapies basées sur l'immunité suite à l'élimination des *MDSC* du micro-environnement tumoral (Trikha and Carson (2014)).

1.5 Quel traitement proposer aux patients atteints de leucémie myélomonocytaire chronique ?

La première attitude du médécin face au diagnostic de leucémie myélomonocytaire chronique est très souvent la mise en place d'une surveillance régulière sans traitement spécifique. Des soins de support sont proposés. Pour corriger une anémie centrale, on propose des injections d'érythropoïétine si le taux circulant d'érythropoïétine est bas ou des transfusions répétées avec chélation du fer. Les infections intercurrentes sont traitées énergiquement. Les patients thrombopéniques reçoivent des transfusions de plaquettes. L'Eltrombopag est actuellement exploré dans les formes très sévèrement thrombopéniques pour tenter de limiter les besoins transfusionnels.

1.5.1 Critères de traitement

Lorsque le nombre de leucocytes augmente et que la rate devient volumineuse, il est habituel de proposer l'Hydrea. Lorsqu'apparaissent des critères de gravité (forte prolifération ou dysplasie sévère), des agents déméthylants (les inhibiteurs de la méthylation de l'ADN (DNMTi, DNA

MethylTransferase inhibitor)) comme l'Azacytidine en Europe et aux USA et la Décitabine aux USA, peuvent être proposés (Treppendahl et al. (2014)) si l'état général du patient le permet. Ces médicaments épigénétiques induisent une réponse objective dans environ 40% des cas. Cette réponse est encore difficile à prévoir, mais une signature épigénétique prédictive vient d'être rapportée (Meldi et al. (2015)).

D'autres médicaments pourraient être envisagés compte-tenu des éléments physiopathologiques évoqués. Du fait de la fréquente hypersensibilité des progéniteurs myéloïdes au *GM-CSF* (Frankel et al. (1998)), il a été proposé de bloquer le récepteur du *GM-CSF* (Padron et al. (2013)). Le défaut d'apoptose des monocytes pourrait aussi être la cible de molécules qui accentuent la dégradation ou préviennent la synthèse des protéines anti-apoptotiques de la famille Bcl2.

Le seul traitement curatif de la leucémie myélomonocytaire chronique est la greffe de cellules souches hématopoïétiques allogéniques. Ce traitement n'est que rarement possible en raison de l'âge des patients. Le risque de réaction du greffon contre l'hôte est très élevé chez les personnes âgées. Cette réaction immunologique qui affecte notamment la peau, le foie et le tube digestif, est responsable du décès d'une part importante des patients chez lesquels la greffe est réalisée.

1.5.2 Agents déméthylants

Les agents déméthylants ont été développés initialement comme des agents cytotoxiques classiques, donc à forte dose. Dans les années 90, les doses d'Azacitidine et Décitabine ont été diminuées de manière à pouvoir prolonger l'exposition des patients aux traitements (Ahuja et al. (2014)). C'est alors qu'une certaine efficacité de ces médicaments a été détectée dans les hémopathies myéloïdes. Il s'est avéré, après quelques années d'expérience, que ces médicaments avaient une efficacité clinique et biologique après plusieurs cycles de traitement et non dès le premier cycle. Il faut attendre 3 à 6 cycles avant d'envisager l'échec. Il s'est avéré également que ces médicaments améliorent la symptomatologie clinique et biologie mais toujours de façon transitoire, la rechute étant systématique.

Dans la leucémie myélomonocytaire chronique, la Décitabine a un effet objectif chez 38% des patients (Braun et al. (2011)) et la majorité des répondeurs rechute dans les 2 ans. Il n'y a pas de thérapeutique standard suite à l'échec de ces traitements et la survie de ces patients est très limitée (Lee et al. (2015)).

L'Azaciditine est injectée en sous-cutanée dans l'avant-bras, la taille ou l'abdomen. Le site d'injection est permuté afin d'éviter les érythèmes. La dose est de $75mg/m^2$ tous les jours pendant 7 (ou 5) jours, suivi d'une période de repos de 21 jours. Les cycles doivent être répétés toutes les 4 semaines et la dose doit être ajustée en fonction de la toxicité. Des retards dans les cycles peuvent être nécessaires afin de laisser à l'organisme le temps de se remettre. La toxicité la plus courante est la myélosuppression, se traduisant par une neutropénie et une thrombocytopénie. Les effets hématologiques indésirables sont le plus souvent observés durant les deux premiers cycles de traitement.

La durée du traitement n'est pas encore définie de manière précise. Chez 91% des patients répondant à l'Azacitidine, la réponse initiale est observée pendant les six premiers cycles. La poursuite du traitement après la réponse initiale améliore la réponse pour près de la moitié des patients. La réponse optimale a été atteinte pour 92% des répondeurs au bout de douze cycles (Silverman et al. (2011)). Similairement, la médiane du temps de réponse à la Décitabine était de plus de trois mois dans plusieurs études (Lübbert et al. (2011) par exemple), et il est recommandé

de continuer le traitement pour au moins quatre cycles, en contrôlant les toxicités. Les traitements doivent être continués tant que le patient montre des bénéfices. Le plus souvent, le traitement est poursuivi si la maladie est stable, non évolutive, même en l'absence de critères objectifs de réponse. La durée médiane de la réponse hématologique est d'environ 13 mois avec l'Azacitidine (Fenaux et al. (2009), Silverman et al. (2002)) et de 9/10 mois avec la Décitabine (Kantarjian et al. (2006), Lübbert et al. (2011)).

Le mécanisme d'action de ces médicaments reste un sujet de controverse. Il existe un effet cytotoxique à forte dose, généralement supérieure à celle accessible pharmacologiquement. Il existe également un effet épigénétique de réduction de la méthylation de l'*ADN* (par inactivation de DNMT1 (Derissen et al. (2013))). Cet effet est supposé inhiber le "silencing" de gènes suppresseurs de tumeur par hyperméthylation des îlots CpG de leur région promotrice et la restauration de fonctions cellulaires normales.

L'Azacitidine est principalement convertie en Azacitine triphosphate, qui est incorporé dans l'*ARN*. Seuls 10 à 20% sont convertis en 5-aza-2′-deoxycytidine triphosphate et sont incorporés dans l'*ADN*. La Décitabine au contraire, convertie en triphosphate, est un déoxyribonucléotide, incorporé dans l'*ADN* (Derissen et al. (2013)). Cette incorporation conduit à la formation d'adduits entre l'*ADN* et DNMT1 (figure 1.3). À forte dose, l'*ADN* n'est plus capable de les éliminer et la cellule meurt. À faible dose en revanche, les adduits formés sont dégradés par le protéasome. La synthèse de l'*ADN* est donc faite en l'absence de DNMT1. De ce fait, la méthylation anormale ne peut pas être reproduite sur le brin fille (McCabe et al. (2009)). Finalement, une faible dose d'Azaciditine ou de Décitabine permet la re-expression de gènes éteints, comme des gènes régulant le cycle cellulaire, ce qui permet de rétablir la différenciation cellulaire, de contrôler la prolifération ou d'accroître l'apoptose des cellules filles (Christman (2002)).

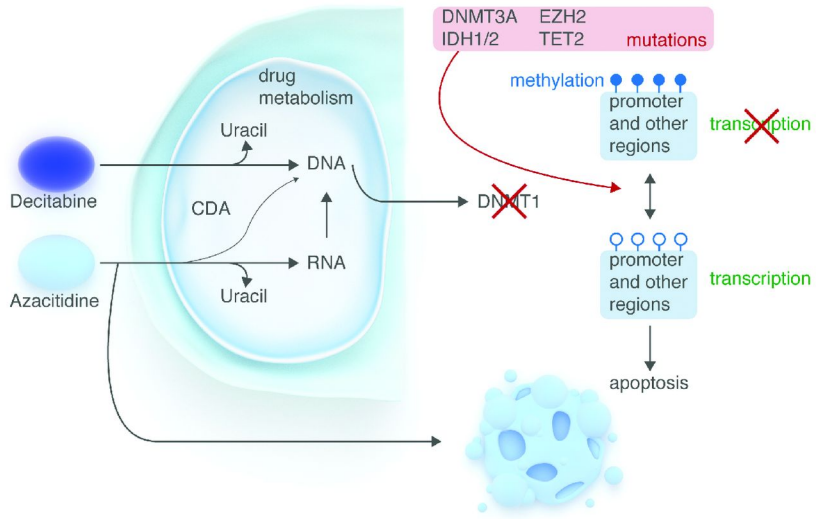


FIGURE 1.3 – *Mode d'action de l'Azacitidine et de la Décitabine (Voso et al. (2014))*

Plusieurs paramètres pharmacologiques peuvent affecter l'efficacité des agents déméthylants (Treppendahl et al. (2014)). Les transporteurs de nucléotides pourraient jouer un rôle dans la mesure où il a été montré *in vitro* que les *DNMTi* les utilisent et que la cytotoxycité dépend de leur présence. Dans la suite du processus d'inhibition de la méthylation de l'*ADN*, il y a monophosphorylation de l'Azacitidine et de la Décitabine par une cytidine kinase et une deoxycytidine kinase, respectivement. L'altération de ces kinases pourrait induire une résistance aux *DNMTi*. Les mutations de la deoxycytidine kinase sont rares chez les patients, mais une sous-expression de ce gène a été observée chez des patients non répondeurs aux agents déméthylants. Enfin,

le niveau d'expression et l'activité enzymatique de la cytidine déaminase semblent influencer la survie des patients traités par *DNMTi*. Cette cytidine déaminase inactiverait la Décitabine et l'Azacitidine par déamination hydrolique irréversible de la kinase cytidine/deoxycytidine en uridine/déoxyuridine. Ainsi, une forte expression de la cytidine déaminase diminue la demi-vie de ces molécules.

Des thérapeutiques de deuxième génération ciblant l'épigénétique sont à l'étude et plusieurs sont engagées dans des essais cliniques. Les trois inhibiteurs les plus prometteurs sont les inhibiteurs de DOT1L, de BET et d'EZH2. La famille BET de bromodomaines lie les histones acétylées au niveau de leurs résidus lysines. DOTL1 est une méthyltransférase de l'histone H3 (Campbell et al. (2014)). Il existe dans les lymphomes malins non Hodgkiniens des mutations hotspots activatrices dans *EZH2* - distinctes des mutations inhibitrices observées dans les hémopathies myéloïdes comme la leucémie myélomonocytaire chronique - qui sont des cibles pour les inhibiteurs d'EZH2 (McCabe et al. (2012)) avec des résultats préliminaires intéressants. Les inhibiteurs d'IDH1 ou d'IDH2 (Wang et al. (2013a)), spécifiques des formes mutées, sont très prometteurs dans les leucémies aiguës myéloïdes avec mutation d'IDH, notamment celles compliquant l'évolution d'une leucémie myélomonocytaire chronique. Ces agents thérapeutiques sont en cours de développement et d'optimisation. Ils devront probablement être associés à d'autres thérapies dans les cancers solides où leur action est très limitée à ce jour.

Altérations moléculaires dans la leucémie myélomonocytaire chronique 2

Les altérations génomiques rapportées dans la leucémie myélomonocytaire chronique sont nombreuses, mais aucune n'est spécifique. Nous décrivons dans ce chapitre les anomalies chromosomiques, mutations somatiques et dérégulations épigénétiques à l'œuvre dans cette pathologie.

2.1 Altérations chromosomiques

Les altérations chromosomiques les plus fréquentes sont les pertes d'hétérozygotie, observées chez près de 1 patient sur 2 (Gondek et al. (2007)). Une perte d'hétérozygotie consiste en la perte du matériel génétique provenant d'un des parents. Le premier mécanisme est la disomie uniparentale acquise. Dans une paire de chromosomes, chacun vient d'un parent différent. Il y a disomie uniparentale lorsque le caryotype est normal (46 chromosomes) mais qu'une ou plusieurs paires de chromosomes ou de fragments de chromosome sont issus d'un seul parent. Les autres mécanismes responsables des pertes d'hétérozygotie sont les délétions ou les pertes chromosomiques, les conversions géniques et les recombinaisons mitotiques (figure 2.1). La perte d'hétérozygotie est un des mécanismes conduisant à la suppression d'un gène suppresseur de tumeur. Le premier allèle est généralement désactivé par mutation somatique, tandis que le deuxième est mis hors service par perte d'hétérozygotie. Ce mécanisme est donc de grande importance dans les maladies génétiques.

La perte d'hétérozogotie la plus fréquente dans la leucémie myélomonocytaire chronique se situe en 4q et englobe systématiquement le gène *TET2* (Jankowska et al. (2009)). Les mutations homozygotes de *CBL* (Grand et al. (2009)) et *TET2* (Mohamedali et al. (2009)) sont associées à des régions de disomies uniparentales.

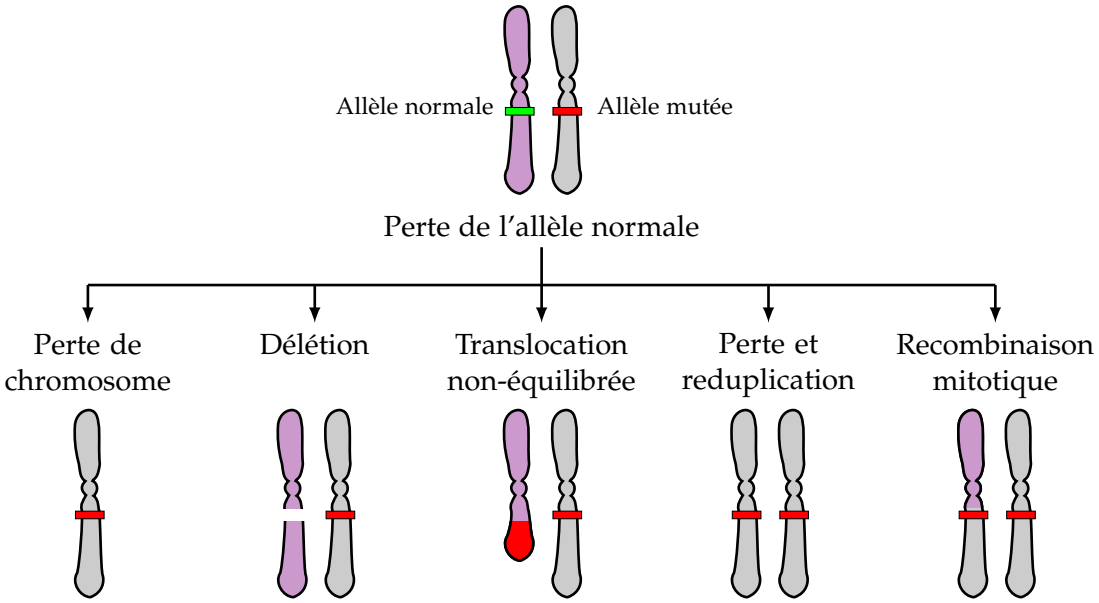


FIGURE 2.1 – *Mécanismes à l'origine des pertes d'hétérozygotie*

Des altérations du nombre de copies surviennent chez environ 30% des patients (Such et al. (2011)). Ces altérations résultent de la perte ou de la duplication partielle ou complète de chromosome. La trisomie 8 est la plus fréquente (environ 7% des cas), suivie de la perte du chromosome Y (environ 4% des cas) et des anomalies du chromosome 7 (environ 1.5% des cas) (Such et al. (2011)). La perte du chromosome Y peut être partielle ou complète et est plutôt associée à un bon pronostic avec les caryotypes normaux et les très rares chromosomes dérivés der(3q). Les autres monosomies et les caryotypes complexes sont de mauvais pronostic. La monosomie 7 a été identifiée dans les années 70 dans les néoplasmes myéloprolifératifs (Petit et al. (1973)), leucémies aiguës myéloïdes (Mitelman et al. (1976)) et leucémies aiguës lymphocytaires (Rowley (1980)). Depuis, elle a été rapportée dans la majorité des leucémies. Dès les années 80, elle a été associée à un mauvais pronostic (Pasquali et al. (1982)). La trisomie 8 semble également associée à un mauvais pronostic (Yunis et al. (1984)). D'autres altérations chromosomiques possibles, quoique rares, sont les translocations réciproques. La plus connue, la t(5 ;12)(q33 ; p13) dont le produit est la protéine de fusion TEL/PDGFβR, est très rare si bien que ces maladies sortent de la définition de la leucémie myélomonocytaire chronique établie par l'*OMS*. Dans le contexte de notre étude, il est intéressant de relever ce travail récemment publié montrant que l'acquisition de nouvelles anomalies cytogénétiques en cours d'évolution, observée chez 20-30% des patients suivis au moins 18 mois, est de mauvais pronostic (Tang et al. (2015)).

2.2 Mutations somatiques

Une trentaine de gènes portant des mutations somatiques de manière récurrente ont été identifiés dans la leucémie myélomonocytaire chronique. L'étude d'Itzykson et al. (2013a) réalisée au sein du laboratoire a montré que 95% des patients présentent au moins une mutation parmi 18 gènes étudiés. Les mutations de ces gènes ont été identifiées par séquençage Sanger de monocytes sanguins $CD14^+$. Cette étude a permis d'estimer la fréquence de mutation de ces 18 gènes chez les patients atteints de leucémie myélomonocytaire chronique (LMMC) (table 2.1) et de conclure qu'aucune mutation ne semble spécifique à la maladie. Ces mutations interviennent dans plusieurs fonctions clés et sont en majorité retrouvées dans les néoplasmes myéloprolifératifs et syndromes myélodysplasiques. Elles ont presque toutes été mises en évidence par une approche "gènes candidats". Les régulateurs épigénétiques sont altérés chez presque tous les pa-

Régulateurs épigénétiques	Facteurs d'épissage	Régulateurs de la signalisation	Facteur de transcription
TET2 : 58%	*SRSF2* : 46%	*NRAS* :11%	*RUNX1* : 15%
ASXL1 : 40%	*ZRSF2* : 8%	*CBL* : 10%	*NPM1* : 1%
IDH2 : 6%	*SF3B1* : 6%	*KRAS* : 8%	*TP53* : 1%
EZH2 : 5%	*U2AF1* : 5%	*JAK2* : 8%	
DNMT3A : 2%		*FLT3* : 3%	
IDH1 : <1%			

TABLE 2.1 – *Fréquence des mutations d'un panel de 18 gènes dans la leucémie myélomonocytaire chronique*

tients, en particulier le gène *TET2* mais aussi le gène *ASXL1*, mutés respectivement chez environ 60 et 40% des patients. Nous trouvons de plus les gènes *IDH2*, *EZH2*, *UTX* et *DNMT3A*, mutés chacun chez moins de 5% des patients. La voie des cohésines a récemment été trouvée altérée aussi (Kon et al. (2013)). Les gènes d'épissage représentent la deuxième catégorie la plus fréquemment affectée dans la leucémie myélomonocytaire chronique avec notamment *SRSF2* muté chez la moitié des patients et les gènes *ZRSF2*, *SF3B1*, *U2AF1* et *LUC7L2* mutés chez moins de 15% des patients. Une autre catégorie est constituée des gènes de la signalisation cytokinique avec *CBL* et *NRAS* mutés chez 10% des patients (Itzykson et al. (2013a)), *JAK2* et *KRAS* mutés chez 8 et 6% des patients (Itzykson et al. (2013a)) ainsi que *SH2B3*, *FLT3*, *KIT*, *MPL*, *LNK*, *RIT1* et des gènes de la voie Notch (<5%). Les facteurs de transcription sont également altérés avec *RUNX1* muté chez environ 15% des patients (Itzykson et al. (2013a)), mais aussi *NPM1*, *BCOR*, *CUX1* et *TP53* (<5%). Enfin, les gènes *SETBP1* et *ETNK1* sont mutés de manière récurrente dans la leucémie myélomonocytaire chronique à une faible fréquence (<5%). Dans les paragraphes suivants, nous répertorions ces gènes en fonction de leur rôle. Certains gènes semblent mutés tout le long de leur partie codante tandis que d'autres sont altérés sur une partie préférentielle (exon ou base) du gène. Le tableau 2.2 fournit les positions préférentiellement mutées dans les gènes principalement mutés de la leucémie myélomonocytaire chronique.

2.2.1 Régulateurs épigénétiques

Les modifications épigénétiques consistent en la modification des histones et la méthylation - déméthylation de l'*ADN*. La figure 2.2 illustre une molécule d'*ADN* compactée autour de nucléosomes subissant des modifications d'histones et de méthylation. La figure 2.3 indique les acteurs de la régulation épigénétique au voisinage d'un nucléosome.

Les gènes *TET2* et *DNMT3A* ont un rôle majeur dans la méthylation de l'*ADN*. La méthylation modifie certaines bases nucléotidiques par ajout d'un groupement méthyle. Chez l'Homme, seul le nucléotide cytosine peut être méthylé au niveau des îlots CpG. La cytosine devient alors une 5-méthylcytosine. La méthylation influence la réparation des mésappariements de l'*ADN* et le niveau d'expression de certains gènes. La relation entre méthylation et expression est complexe. De manière générale, une faible méthylation implique la transcription d'un gène en *ARN* messager (ARNm) tandis qu'une forte méthylation la réprime. Par exemple, lorsque le promoteur d'un gène est méthylé, la transcription de ce gène en *ARNm* ne peut plus s'effectuer. La méthylation de l'*ADN* a donc un rôle clé dans l'expression des gènes. La méthylation est un processus réversible mais les mécanismes de déméthylation ne sont pas bien connus.

TET2 (*ten-eleven-translocation 2*) est le gène le plus fréquemment muté dans la leucémie myélomonocytaire chronique, à hauteur de 60% environ. Les mutations de *TET2* (Delhommeau et al. (2009)), principalement des pertes de fonction, sont présentes dans l'ensemble des hémopathies myéloides, à une fréquence de 11%, 20%, 19% et 37% dans les néoplasmes myéloprolifératifs,

Gène	Position	Localisations des mutations
TET2	4q24	dispersées
ASXL1	20q11	exon 12
DNMT3A	2p23	R882 et dispersées
IDH1	2q33	R132
IDH2	15q26	R140, R172
EZH2	7q35	dispersées
SRSF2	17q25	P95
U2AF1	21q22	S34, R156, Q157
ZRSF2	Xp22	dispersées
SF3B1	2q33	K200, K667
LUC7L2	7q34	dispersées
SH2B3	12q24	dispersées
NRAS	1p13	exons 2 (G12) et 3
KRAS	12p12	exons 2 (G12) et 3
JAK2	9p24	V617F
CBL	11q23	exons 8 et 9
RUNX1	21q22	dispersées
CUX1	7q22.1	dispersées
BCOR	Xp11.4	dispersées
RIT1	1q21	dispersées
STAG2	Xq25	dispersées
ETNK1	12p12	N244S, H243Y

TABLE 2.2 – *Localisations des mutations somatiques dans les gènes les plus fréquemment mutés dans la leucémie myélomonocytaire chronique*

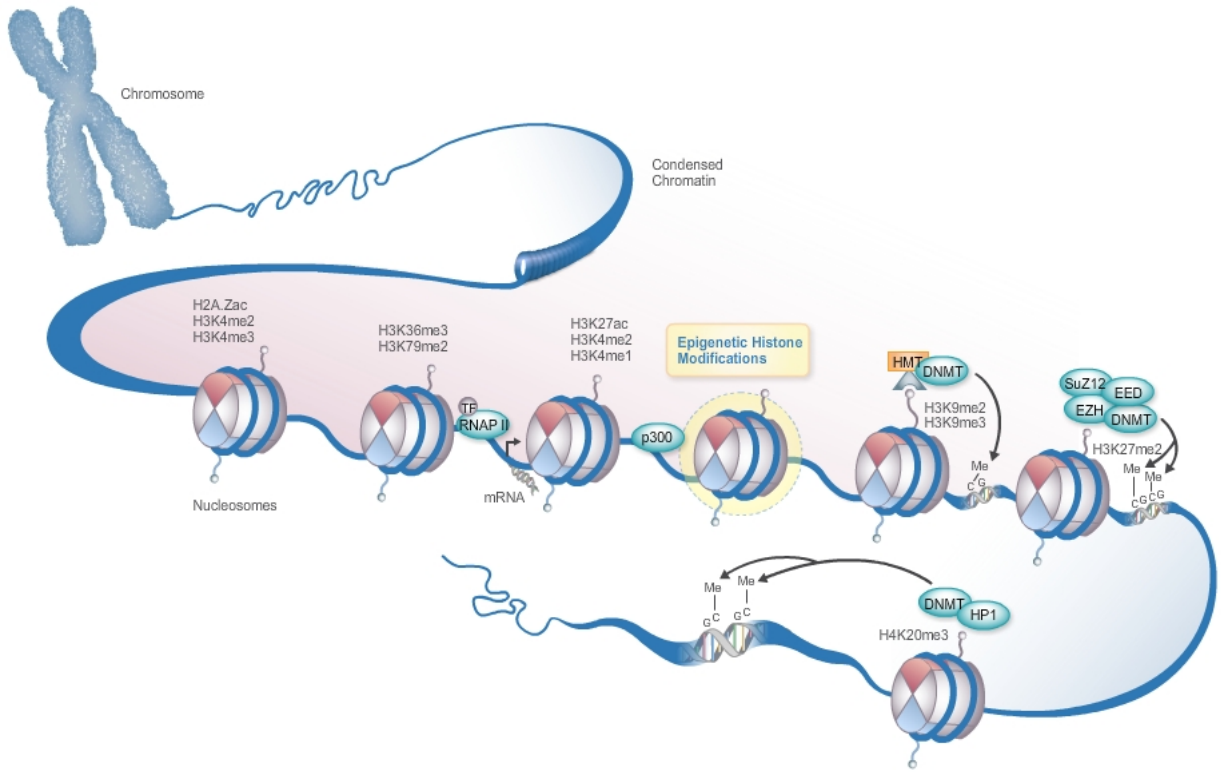


FIGURE 2.2 – *Modifications épigénétiques se produisant le long d'une molécule d'ADN (http ://www.abcam.com/)*

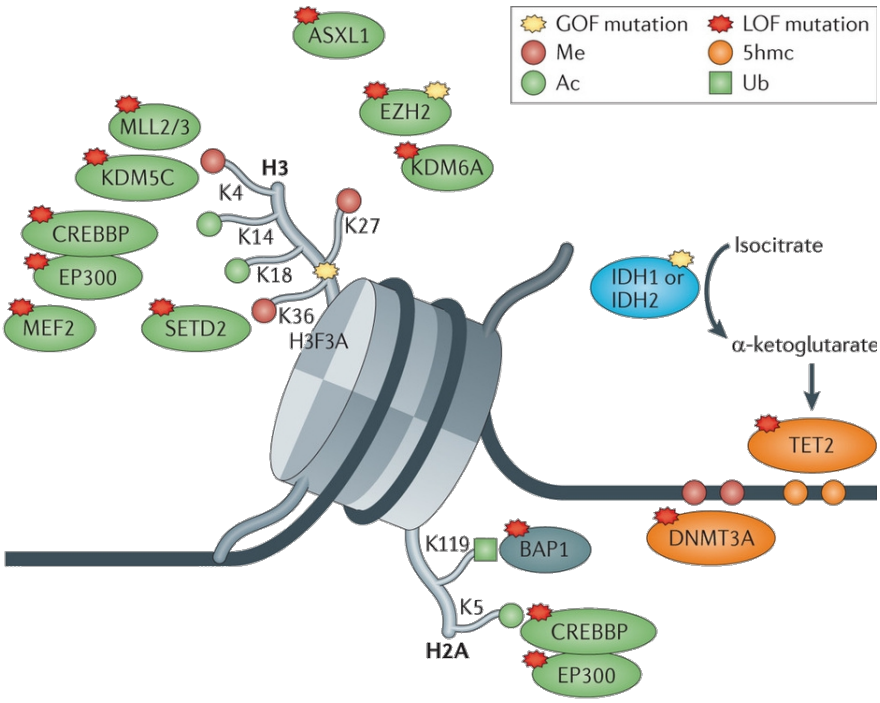


FIGURE 2.3 – *Principaux acteurs de la régulation épigénétique (Watson et al. (2013))*

syndromes myélodysplasiques, leucémies aiguës myéloïdes et syndromes frontières respectivement. Ce gène fait partie des protéines de la famille TET (TET1, 2 et 3). Ces protéines partagent deux régions très conservées : une au centre de la protéine et l'autre à l'extrémité C-terminale. Ces deux domaines forment le domaine catalytique de *TET2* (Hu et al. (2013)). La région centrale est un domaine riche en cystéines. Il joue un rôle important dans l'interaction TET2-*ADN* et dans l'activité enzymatique des TET. Toute son importance réside dans sa structure puisque ce domaine est essentiel pour assurer la structure globale de la protéine. La région terminale présente la structure en double chaîne d'hélice β des oxygénases dépendantes du Fe(II) et de l'oxoglutarate qui catalyse la conversion de la 5-methylcytosine (5mc) dans l'*ADN* en 5-hydroxymethylcytosine (5-hmC) (Ko et al. (2010)). Les 5hmc initient la déméthylation de l'*ADN*. En effet, les enzymes de la famille TET sont impliquées dans la déméthylation par un mécanisme d'hydroxylation, suivi d'une première étape de déamination par des cytidines déaminases (AID/APOBEC), et d'une deuxième étape faisant intervenir des enzymes impliquées dans les mécanismes de réparation par excision de base (Guo et al. (2011)).

DNMT3A (*DNA methyltransferase 3A*) fait partie de la famille DNMT (ADN méthyltransférase). Il existe trois familles d'ADN méthyltransférases. DNMT3, qui regroupe les gènes *DNMT3A* et DNMT3B, intervient dans la méthylation *de novo*. *DNMT3A* intervient dans la méthylation des séquences régulatrices de l'expression des gènes. Ses mutations, décrites par Jankowska et al. (2011), entraînent des pertes de fonction. DNMT3B est impliqué dans la méthylation des séquences d'*ADN* satellite (qui sont des séquences d'*ADN* répétées) et des centromères. *DNMT1* gère le maintien des profils de méthylation au cours des divisions cellulaires. Le rôle de *DNMT2* n'est actuellement pas connu.

Les histones sont des protéines riches en acides aminés basiques de faible masse moléculaire. Leur domaine C-terminal est très conservé, depuis les archées. Les histones sont présentes dans le

noyau des cellules eucaryotes. Ce sont les principaux composants protéiques des chromosomes. Leur rôle est de compacter l'*ADN* autour du nucléosome : l'*ADN* est enroulé autour des histones par interaction entre la charge positive des histones basiques et les groupements phosphates de l'*ADN* qui portent des charges négatives. Le nucléosome est formé des quatre paires d'histones H2a, H2b, H3 et H4. Cette compaction de l'*ADN* par les histones est impliquée dans plusieurs processus qui ont lieu dans le noyau comme la transcription, la réplication ou la réparation en contrôlant l'accès à l'*ADN*. Les modifications des histones comprennent la méthylation, l'acétylation, l'ubiquitinylation et la phosphorylation. Nous décrivons dans la suite les fonctions clés des gènes *ASXL1*, *EZH2* et *UTX* impliqués dans la modification des histones. Les mutations d'*UTX* sont très rares, de l'ordre de 1%.

ASXL1 (*Additional sex combs-like protein 1*) est le troisième gène le plus souvent muté dans la leucémie myélomonocytaire chronique. Ses mutations sont retrouvées chez 35-40% des patients (Itzykson et al. (2013a)). Les mutations d'*ASXL1*, pertes de fonction, ont été décrites par Gelsi-Boyer et al. (2009) dans les syndromes myélodysplasiques et la leucémie myélomonocytaire chronique. L'étude d'Itzykson et al. (2013a) a montré que les mutations d'*ASXL1* sont de mauvais pronostic pour une survie globale et survie sans transformation en leucémie aiguë myéloïde pour les patients *LMMC*. Ce résultat a été montré dans d'autres pathologies (Devillier et al. (2015), Chen et al. (2014)). *ASXL1* est un membre du groupe de protéines "Polycomb". Ces protéines sont nécessaires à la maintenance de la répression des gènes homéotiques. ASXL1 semble "interrompre" la chromatine à certaines positions spécifiques, permettant d'une part l'activation de la transcription de certains gènes et d'autre part la répression d'autres gènes.

EZH2 (*Enhancer of zeste homolog 2*) fait également partie du groupe "Polycomb". C'est la sous-unité catalytique du complexe PRC2/EED-EZH2, qui méthyle les résidus 'Lys-9' (H3K9me) et 'Lys-27' (H3K27me) de l'histone H3, conduisant à la répression transcriptionnelle du gène cible. Les mutations pertes de fonction d'*EZH2* ont été rapportées par Jankowska et al. (2011).

UTX (*Histone demethylase, Ubiquitously-transcribed X chromosome tetratricopeptide repeat protein*) déméthyle spécifiquement le résidu 'Lys-27' de l'histone H3, jouant ainsi un rôle central dans le code des histones. *UTX* régule également le développement de la région postérieure de l'embryon en régulant l'expression des gènes HOX. Les mutations d'*UTX* ont été rapportées par Jankowska et al. (2011).

Les phénomènes de méthylation de l'*ADN* et de modification des histones sont des événements corrélés. En effet, la méthylation de l'*ADN* implique la modification des histones et *vice-versa*. Certains gènes ont un rôle à la fois dans la méthylation de l'*ADN* et dans la méthylation des histones, comme *IDH1* et *IDH2* et les composants des complexes cohésines *STAG2*, *SMC3*, *SMC1A* et *RAD21*.

IDH1 (*isocitrate dehydrogenase 1*) et ***IDH2*** (*isocitrate dehydrogenase 2*) sont des enzymes isocitrates déhydrogénases qui se localisent dans le cytoplasme et la mitochondrie respectivement. Ces enzymes lient le NAD (Nicotinamide adenine dinucleotide), ont une activité isocitrate déhydroxygénase (NAD+) et lient les ions magnésium. Dans les cellules humaines, la production et l'utilisation d'α-ketoglutarate (α-KG) sont assurées par quatre enzymes : *IDH1*, *IDH2*, IDH3 et GDH. L'α-KG est utilisée pour le cycle de Krebs, les réactions anaplérotiques, la synthèse d'acides gras et l'hydroxylation acide de protéines et acides nucléiques (figure 2.4).

Les mutations d'*IDH1* et *IDH2* sont principalement retrouvées dans les gliomes (Yan et al. (2009)) et les hémopathies myéloïdes (leucémies aiguës myéloïdes (Marcucci et al. (2010)) et néoplasmes myéloprolifératifs (Pardanani et al. (2010), Tefferi et al. (2010))). Ces mutations inhibent la

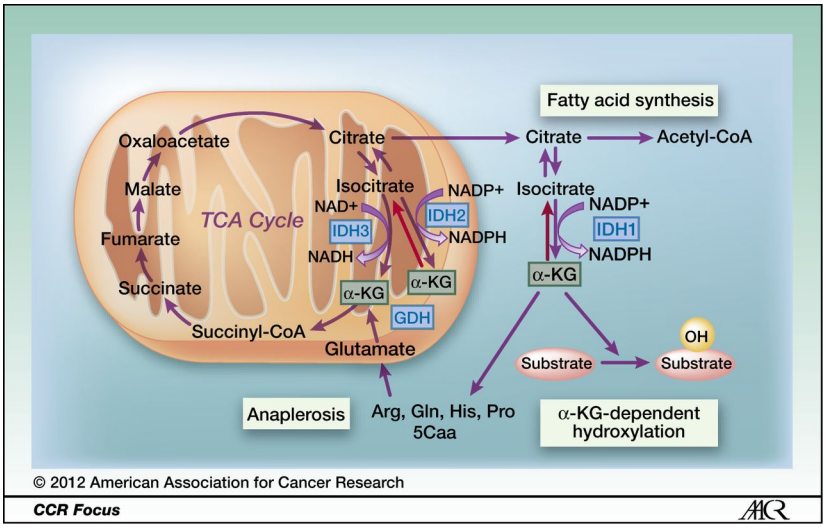


FIGURE 2.4 – *Rôle d'IDH1/2 dans la production et l'utilisation d'α-KG (Yang et al. (2012))*

déméthylation des histones et de l'*ADN* et altèrent la régulation épigénétique. Les mutants gain de fonction d'*IDH1/2* induisent la perte de leur fonction de production d'*α*-KG et l'acquisition d'une nouvelle activité conduisant à la production de 2-hydroxyglutarate (2-HG). Le dosage de 2-HG est un biomarqueur traduisant la présence d'une mutation d'IDH et mesurant la maladie résiduelle dans le suivi des leucémies aiguës myéloïdes avec mutation d'IDH (Janin et al. (2014)). Les mutations des gènes IDH conduisent à l'inhibition des déméthylases lysine-spécifique et de la famille TET (figure 2.5). Ces inhibitions altèrent le contrôle épigénétique de la différenciation des cellules souches et des progéniteurs (Yang et al. (2012)).

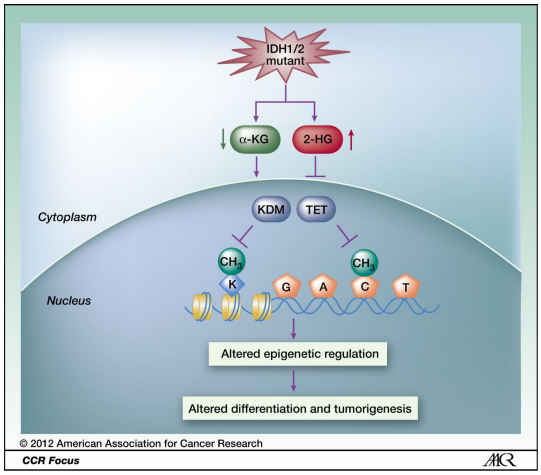


FIGURE 2.5 – *Anomalies liées aux mutations d'IDH1/2 (Yang et al. (2012))*

Les cohésines ont pour principal objectif le maintien des deux chromatides sœurs après réplication. Le complexe cohésine semble former un anneau protéique à l'intérieur duquel les chromatides sœurs peuvent être piégées. À l'anaphase, le complexe est clivé et se dissocie de la chromatine, permettant aux chromatides sœurs de se séparer. Ce complexe pourrait jouer un rôle dans l'ascension polaire durant la mitose. Les cohésines sont également nécessaires à la réparation de l'*ADN* et la régulation d'expression génique des cellules en prolifération et post-mitotique. Le complexe formé regroupe quatre sous-unités (SMC1, SMC3, RAD21 et STAG) (figure 2.6) et des molécules régulatrices (PDS5, NIPBL et ESCO). Dans la leucémie myélomonocytaire chronique,

des mutations somatiques de *STAG2* uniquement ont été détectées, alors que dans les néoplasmes myéloïdes, les quatre sous-unités peuvent être mutées (Kon et al. (2013)).

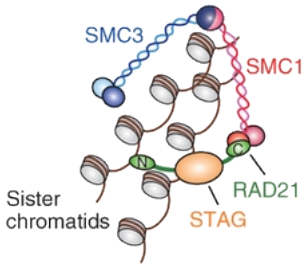


FIGURE 2.6 – *Les sous-unités du complexe cohésine (Kon et al. (2013))*

2.2.2 Facteurs d'épissage

Chez les eucaryotes, l'épissage est le processus par lequel les *ARN* transcrits peuvent subir des étapes de coupure et ligature qui conduisent à la suppression de certaines régions dans l'ARN mature. Les introns de l'ARN pré-messager sont excisés, certains exons peuvent l'être également (figure 2.7) par le splicéosome. Les *ARNm* matures donnent plusieurs isoformes, qui peuvent résulter en des protéines différentes. L'épissage est assuré par le splicéosome, un complexe de protéines (figure 2.8) qui assemblent les sites 5' et 3' marquant la jonction intron-exon. Les acteurs assurant l'épissage sont schématisés figure 2.8. L'assemblage du splicéosome et son réarrangement suivent un ordre très conservé (de la levure aux mammifères) avec la formation des complexes E, A, B1, B2, C1 et C2. *SRSF2*, *U2AF1*, *SF3B1* et *ZRSR2* sont mutés chez 85% des patients avec syndromes myélodysplasiques, de manière mutuellement exclusive. Les mutations de *SRSF2*, *U2AF1*, *SF3B1* sont des pertes de fonction, délétion et silencing du promoteur.

SRSF2 (*Serine/arginine-rich splicing factor 2*) est muté chez 1 patient sur 2 environ, presque exclusivement sur le codon P95 (Meggendorfer et al. (2012), Itzykson et al. (2013a)). Il est nécessaire à l'épissage de l'ARN pré-messager puisqu'il participe aux étapes précoces de l'épissage en interagissant avec les composants du splicéosome pendant son assemblage. *SRSF2* est également requis pour les interactions des snRNPs (small nuclear ribonucleoproteins) avec l'ARN pré-messager. ***U2AF1* ou U2AF35** (*U2 small nuclear RNA auxiliary factor 1*, *U2 auxiliary factor 35 kDa subunit*) joue un rôle essentiel dans l'épissage constitutif et enhancer-dependant (dépendant des séquences "enhancer" situées dans l'exon : Exonic Splicing Enhancer (ESE)) en médiant les interactions protéines-protéines et *ARN*-protéines nécessaires à la sélection précise du site 3' d'épissage. *U2AF1* est muté chez environ 10% des patients (Itzykson et al. (2013a)). ***SF3B1*** (*Splicing factor 3B subunit 1*) participe à l'assemblage des complexes A et E. Il est muté chez environ 10% des patients (Itzykson et al. (2013a)). *SF3B1* est une sous-unité de facteur d'épissage SF3B nécessaire à l'assemblage du complexe A formé par la liaison stable entre le snRNP U2 et la séquence de branchement de l'ARN pré-messager. La liaison, séquence indépendante du complexe SF3A/SF3B avant le site de branchement est essentielle car pourrait ancrer U2 à l'ARN pré-messager. **ZRSR2** (*U2 small nuclear ribonucleoprotein auxiliary factor 35 kDa subunit-related protein 2*) est muté chez environ 8% des patients (Itzykson et al. (2013a)). Il intervient dans l'épissage de certains types d'introns et dans l'assemblage du pré-splicéosome. ***LUC7L2*** a été identifié comme jouant un rôle dans l'épissage mais sa fonction précise n'est pas connue. Les mutations dans ce gène ont été décelées en 2013 (Singh et al. (2013)) dans les hémopathies myéloïdes.

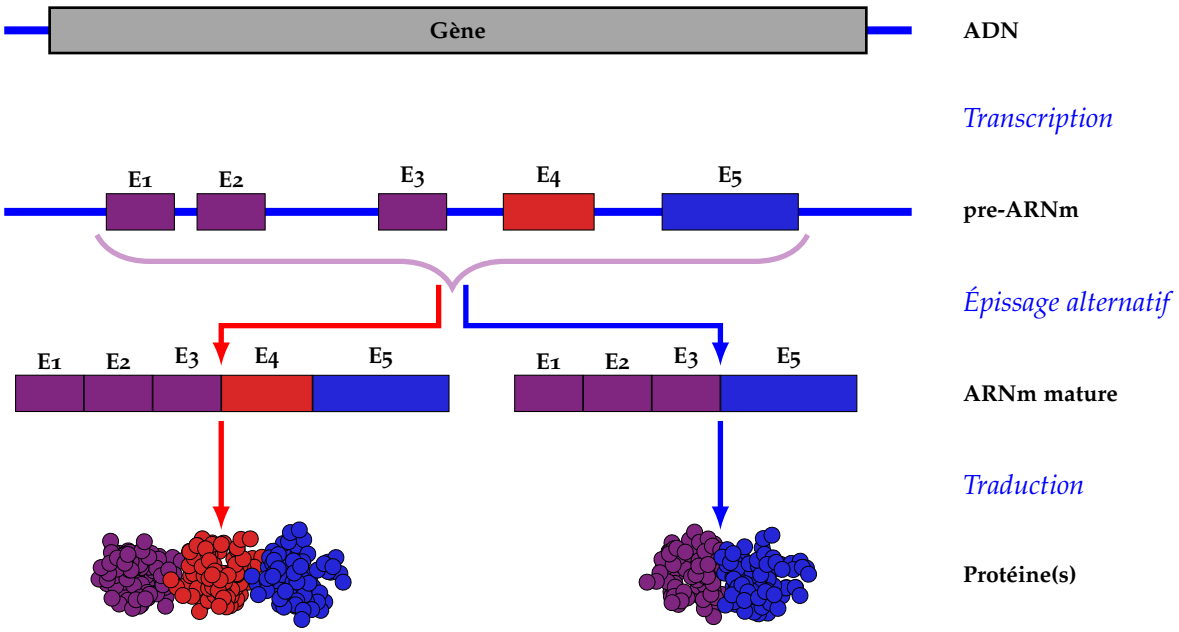


FIGURE 2.7 – *Épissage et Épissage alternatif*

2.2.3 Régulateurs de signalisation cytokinique

La signalisation cytokinique est un processus qui permet à une cellule d'adapter son comportement aux besoins de l'organisme. La réponse est soit une déformation du cytosquelette, soit une modification d'expression génique ou encore un changement métabolique. La signalisation est caractérisée par un récepteur et son ligand, une cytokine le plus souvent, des transducteurs permettant de diffuser l'information dans le cytoplasme jusqu'à l'effecteur qui est souvent un facteur de transcription et enfin, des régulateurs qui modulent la signalisation.

CBL (*Casitas B-lineage lymphoma proto-oncogene*) est un gène suppresseur de tumeur dont les mutations gain de fonction sont souvent associées à une disomie uniparentale acquise en 11q dans les hémopathies myéloïdes (Sanada et al. (2009)). Il s'agit d'une protéine adaptatrice qui fonctionne comme un régulateur négatif de plusieurs voies de signalisation qui sont déclenchées par activation de récepteurs de surface cellulaire. *CBL* est une protéine ubiquitine ligase E3, qui permet la dégradation de ses substrats par le protéasome. *CBL* reconnait certains récepteurs à activité tyrosine kinase, comme KIT, FLT1, PDGFRα, PDGFRβ, EGFR et CSF1R lorsqu'ils sont actifs.

***KRAS* (*GTPase KRas*) et *NRAS* (*Transforming protein N-Ras, GTPase NRas*)** sont des protéines RAS à activité GTPase intrinsèque (lient GDP (Guanosine Di Phosphate) et GTP (Guanosine Tri Phosphate)). Ce sont des transducteurs du signal qui participent à l'activation de la voie MAPK. Ces proto-oncogènes favorisent une prolifération cellulaire non contrôlée lorsqu'ils sont mutés. Leurs mutations, souvent hotspots, ont été initialement détectées dans les leucémies aiguës myéloïdes (Bos et al. (1985)), puis dans les cancers colorectaux (Bos et al. (1987)) et syndromes myélodysplasiques (Hirai et al. (1987), Liu et al. (1987)). Leurs mutations conduisent à un gain de fonction et sont répertoriées dans de nombreux cancers : de 90% environ dans les cancers colorectaux à environ 10% pour *NRAS* et pour *KRAS* dans la leucémie myélomonocytaire chronique.

JAK2 (*Janus kinase 2*) a un rôle crucial dans la transduction du signal *via* ses associations avec les récepteurs de type 1 comme les hormones de croissance, la prolactine, la leptine, l'érythropoïétine, la thrombopoïétine et ses récepteurs de type 2, comme IFN$_\alpha$, IFN$_\beta$, IFN$_\gamma$ et plusieurs

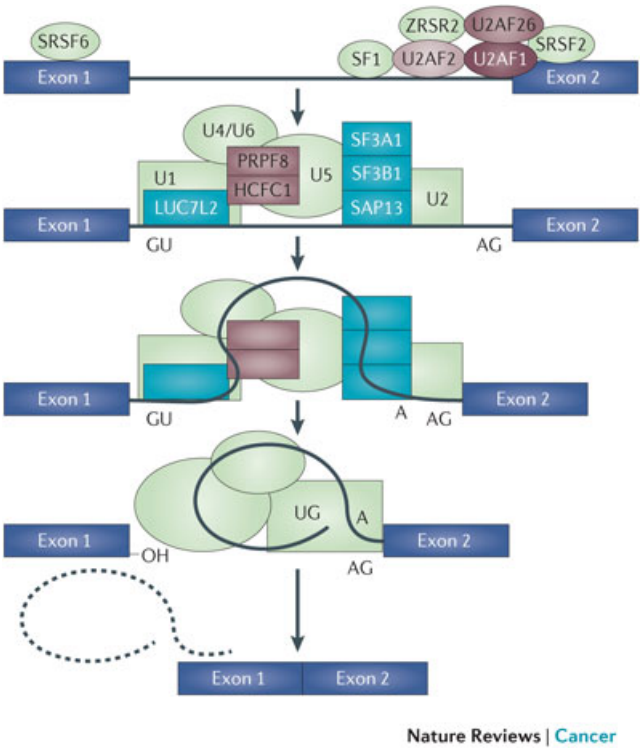


FIGURE 2.8 – *Complexe de protéines assurant l'épissage (Raza and Galili (2012))*

interleukines. *JAK2* peut également phosphoryler les membres de la famille STAT (Signal Transducer and Activator of Transcription). Les STATs phosphorylés peuvent se dimériser et entrer dans le noyau. En tant qu'activateurs de la transcription, ils vont modifier l'expression génique de leurs cibles. *JAK2* est une tyrosine kinase impliquée dans différents processus comme la croissance, le développement, la différenciation ou la modification des histones. *JAK2* intervient dans d'importants événements de signalisation de l'immunité innée et acquise. Ses mutations gains de fonction (James et al. (2005), Levine et al. (2005)) semblent être des événements secondairement acquis dans la leucémie myélomonocytaire chronique, contrairement au cas des néoplasmes myéloprolifératifs. De plus, contrairement à la polyglobulie de Vaquez et à la thrombocytémie essentielle où la fréquence de mutation est d'environ 90% et 50% respectivement, les mutations de *JAK2* dans la leucémie myélomonocytaire chronique ne sont observées que dans 5% des cas environ (Itzykson et al. (2013a)).

SH2B3 ou LNK (*SH2B adapter protein 3*, *Lymphocyte adapter protein*) transmet le signal d'activation du récepteur des lymphocytes T aux transducteurs phospholipase C-γ-1, GRB2 (voie MAPK) et phosphatidylinositol 3-kinase (voie PI3K/AKT). C'est un régulateur négatif de la signalisation de JAK2. Ses mutations, initialement décrites par Oh et al. (2010), entraînent des pertes de fonction.

FLT3 (*Fms-like tyrosine kinase 3*) est un récepteur à activité tyrosine kinase pour son ligand, la cytokine FLT3LG. Il régule la différenciation, la prolifération et la survie des progéniteurs hématopoïétiques et des cellules dendritiques. *FLT3* peut activer les voies MAPK et AKT. Les mutations, causant une activité kinase constitutive, favorisent la prolifération cellulaire et la résistance à l'apoptose *via* l'activation de plusieurs voies de signalisation. ***KIT*** (*Mast/stem cell growth factor receptor*) est une protéine tyrosine kinase qui joue un rôle important dans la régulation de survie et prolifération cellulaire, l'hématopoïèse, la maintenance du compartiment de cellules souches, la gamétogénèse, la migration et la mélanogénèse. *KIT* peut activer plusieurs voies de signalisation, dont les MAPK/ERK. Ses mutations (Lorenzo et al. (2006)) sont présentes chez moins de 5% des

patients *LMMC*, comme les mutations de *RIT1* (Gomez-Segui et al. (2013)). ***RIT1*** (*Ras-like without CAAX protein 1, GTP-binding protein*) est un membre de la famille RAS. La protéine encodée est impliquée dans la régulation de cascades de signalisation dépendant de la MAPK p38, activées en réponse au stress cellulaire. Cette protéine intervient dans le développement et la regénération des neurones.

2.2.4 Facteurs de transcription

La transcription d'un gène en une molécule d'*ARNm* fait intervenir un ensemble protéique complexe, incluant l'ARN polymérase. La première étape de la transcription est la reconnaissance du gène à transcrire. Une séquence particulière de nucléotides indique le début du gène dans le promoteur. Une protéine spécifique du gène à transcrire s'y fixe. Ce complexe lit la molécule d'*ADN*. Elle déroule d'abord la molécule d'*ADN*, puis sépare les deux brins et assemble les bases azotées en se servant du brin complémentaire comme matrice pour aboutir à la molécule d'*ARN*. Derrière elle, les deux brins se rassemblent. Quand l'ARN polymérase rencontre le site de terminaison du gène, elle se sépare de l'*ADN* et l'*ARN* est libéré de la chaîne d'*ADN*. L'*ARN* transcrit, ARN pré-messager, subit différentes modifications, avant d'être traduit. L'épissage par exemple suit directement la transcription (de la Mata et al. (2003)). L'*ARN* synthétisé donne alors l'*ARN* mature utilisé pour la traduction en complexes protéiques.

RUNX1 ou AML1 (*Runt-related transcription factor 1, acute myeloid leukemia 1 protein* (AML1)) est le facteur de transcription le plus souvent muté dans la leucémie myélomonocytaire chronique, chez environ 15% des patients. Ses mutations pertes de fonction ont été décrites par Kuo et al. (2009). Les mutations acquises de *RUNX1* ou les protéines de fusion impliquant *RUNX1* sont les anomalies génétiques les plus fréquentes dans les leucémies. *RUNX1* est considéré comme le facteur de transcription clef dans l'hématopoïèse définitive. Il régule la différenciation des *CSH* en cellules matures.

Les autres facteurs de transcription sont mutés chez moins de 5% des patients. ***BCOR*** (*BCL-6 corepressor*) est un corépresseur transcriptionnel. Il pourrait inhiber spécifiquement l'expression génique lorsqu'il est recruté au niveau des régions promotrices par des protéines liant des séquences spécifiques de l'*ADN* telles que BCL6 et MLLT3. Cette répression pourrait être médiée au moins en partie par le recrutement de HDAC (Histone DeACetylase) qui désacétyle les histones. Ses mutations somatiques ont été rapportées par Damm et al. (2013) dans les syndromes myélodysplasiques. ***CUX1*** (*Homeobox protein cut-like 1*) est un membre de la famille des protéines à homéodomaine. Il pourrait réguler l'expression génique, la morphogénèse et la différenciation. *CUX1* a été mis en évidence comme *TET2* par intersection de régions perdues sur le chromosome 7 de patients atteints de néoplasmes myéloprolifératifs (Thoennissen et al. (2011)). Il pourrait également jouer un rôle dans la progression du cycle cellulaire. ***NPM1*** (Nucleophosmin) a de nombreuses fonctions, entre autres chaperone, biosynthèse et transport des ribosomes, stabilité génomique et duplication du centrosome. Ses mutations sont très peu fréquentes dans la leucémie myélomonocytaire chronique, de l'ordre de 1% (Itzykson et al. (2013a)). Il existe également des mutations de ***TP53***, relativement fréquentes dans les syndromes myélodysplasiques et de très mauvais pronostic. Elles sont très rares dans la leucémie myélomonocytaire chronique, classiquement moins de 1% des patients.

Enfin, ***ETNK1*** (*Ethanolamine kinase 1*) ne rentre pas dans les catégories précédentes. Il code une éthanolamine kinase qui catalyse la première étape de la voie de biosynthèse phosphatidyléthanolamine *de novo*. La phosphatidyléthanolamine participe à plusieurs processus biochimiques

comme la définition de l'architecture membranaire, la progression de la cytokinèse pendant la division cellulaire ou l'activité des complexes respiratoires dans la membrane interne de la mitochondrie. Les mutations d'*ETNK1* ont été trouvées dans une région très conservée à l'intérieur du domaine kinase et pourraient endommager l'activité catalytique de l'enzyme. Ces mutations pertes de fonction ont été décelées chez environ 3% des patients *LMMC* et chez environ 9% des patients atteints de leucémie myéloïde chronique atypique (Gambacorti-Passerini et al. (2015)).

2.3 Architecture du clone leucémique

Itzykson et al. (2013b) ont étudié le niveau et l'ordre d'apparition des mutations dans 18 gènes au cours de l'hématopoïèse, en étudiant les stades les plus immatures chez 28 patients, soit dans la moelle osseuse (*CSH*, *MPP*, *CMP* et *GMP*), soit dans le sang ($CD34^+/CD38^-$), à l'échelle unicellulaire. Chez les 28 patients, au moins une mutation a été trouvée dans plus de 75% des *CSH*, marquant une apparition précoce des mutations, dès le stade de la *CSH* ou d'un progéniteur très immature. En étudiant les *CSH* des 19 patients porteurs d'au moins 2 mutations somatiques, une accumulation linéaire des mutations a été observée chez la plupart des patients. L'ordre d'acquisition des mutations n'est pas fixe mais le plus souvent, la première mutation apparaît dans un régulateur épigénétique (le plus souvent *TET2* ou *ASXL1*) et une deuxième mutation affecte les facteurs d'épissage ou les régulateurs de la signalisation. Souvent, les mutations de *TET2* associées ou non à des mutations de *SRSF2* ou *ASXL1*, précèdent les mutations de la voie Ras. Une architecture clonale complexe peut être observée à la suite de recombinaisons mitotiques, engendrant des sous-clones minoritaires. De plus, l'étude montre que les mutations de *TET2*, *IDH2* et *IDH1* semblent mutuellement exclusives alors que certaines combinaisons de mutations sont fréquentes, comme *TET2* et *SRSF2* ou *ASXL1* et *SRSF2*. Dans 14 des 19 patients, la première mutation est présente dans plus de 85% des *CSH*, *MPP*, *CMP* et *GMP* alors que la deuxième mutation était plus fréquente dans les *GMP* que dans les *CSH* et *MPP*, suggérant que ces événements secondaires donnent un avantage au clone pendant la différenciation myéloïde.

2.4 Niveau d'expression génique dans les cellules leucémiques

La méthylation de l'*ADN* est un mécanisme épigénétique, *i.e.* induisant un changement d'expression génique sans altération de la séquence d'*ADN*. La méthylation de l'*ADN* se produit essentiellement aux positions 5' des cytosines dans des régions du génome riche en CpG di-nucléotides, appelés îlots CpG, résultant en la sous-expression ou extinction de gènes. Ce processus est indispensable au développement normal. La méthylation est assurée par DNMT1, qui maintient le profil de méthylation au cours des réplications et par DNMT3A et B qui méthylent les îlots CpG non méthylés. Les membres de la famille TET au contraire suppriment les groupements méthyles au niveau des îlots CpG. Ce mécanisme nécessite de l'α-ketoglutarate dont la synthèse implique IDH1 et IDH2. L'altération du niveau de méthylation dans le génome suite à un traitement est décrite dans la première partie des résultats chapitre 5.

Il existe une hypermethylation globale de l'*ADN* dans les cellules leucémiques. Des altérations du processus de méthylation / déméthylation des cytosines de l'*ADN* peuvent altérer l'expression des gènes dans une cellule. La méthylation excessive du promoteur du gène *TIF1-γ* (Aucagne et al. (2011)) a été détectée dans les cellules de 35% des patients *LMMC*. Les souris invalidées pour *TIF1-γ* développent un phénotype proche de celui de la leucémie myélomonocytaire chronique. Ces résultats suggèrent que des anomalies épigénétiques participent à l'émergence du clone leucémique.

D'autres altérations de l'expression des gènes dans les monocytes de patients ont été détectées

par une approche plus systématique. Le niveau d'expression de certains gènes comme *CMYB* et *CJUN* (Braun et al. (2011)) pourrait influencer la réponse aux agents déméthylants. Les mécanismes moléculaires à l'origine de l'expression anormale de certains gènes n'ont pas encore été élucidés. Plusieurs facteurs autres que le niveau de méthylation des cytosines au niveau des îlots CpG contribuent à la dérégulation de l'expression génique.

L'altération des gènes codant des régulateurs épigénétiques comme *ASXL1*, *EZH2* ou *UTX* a un rôle parfois encore controversé. Lorsque les mutations de ces régulateurs sont pertes de fonction, ils entraîneraient la sous-expression de leurs partenaires. Les mutations des facteurs d'épissage modifient la qualité ou la quantité des isoformes d'un gène. La mutation de *SRSF2* entraîne la sous-expression d'EZH2 en altérant son épissage (Kim et al. (2015)), démontrant un lien entre épissage et épigénétique. Les altérations des facteurs de transcription ont aussi un impact sur l'expression des gènes, par exemple, les mutations de *RUNX1* perturbent la mégacaryopoïèse. Enfin, les régulateurs de la signalisation agissent directement ou indirectement sur l'expression de gènes contrôlant la mort cellulaire comme *MCL1* et *BCLXL*, la sénescence comme P21, ou la réponse précoce *FOX* et *CJUN* suite à la mutation de *KRAS*.

2.5 Anomalies d'épissage

L'épissage aberrant d'un gène résulte le plus souvent d'une mutation au niveau de son site d'épissage ou d'une anomalie de la machinerie d'épissage par mutation d'un facteur d'épissage. La mutation d'un facteur de transcription peut également induire un épissage alternatif.

Différentes anomalies d'épissage ont été répertoriées : les inclusions ou exclusions d'un ou plusieurs exons, les sites d'épissage alternatif en 3', les sites d'épissage alternatif en 5', les rétentions d'introns, les exons mutuellement exclusifs ainsi que les promoteurs ou sites polyA multiples (figure 2.9). L'étude des anomalies d'épissage dans la leucémie myélomonocytaire chronique est présentée chapitre 6.

Les anomalies d'épissage caractéristiques de la leucémie myélomonocytaire chronique ne sont pas encore clairement identifiées. On sait que la mutation de *SRSF2* entraîne un épissage alternatif d'EZH2. La mutation d'*U2AF1* entraîne des anomalies de l'épissage de 35 gènes du cycle cellulaire et du processing d'*ARN* (Przychodzen et al. (2013). La méthylation de l'*ADN* (DNMT3B), l'inactivation du chromosome X (H2AFY), la réponse aux dommages à l'*ADN* (ATR, FANCA) ou encore l'apoptose (CASP8) sont des cibles d'U2AF1 (Ilagan et al. (2015)).

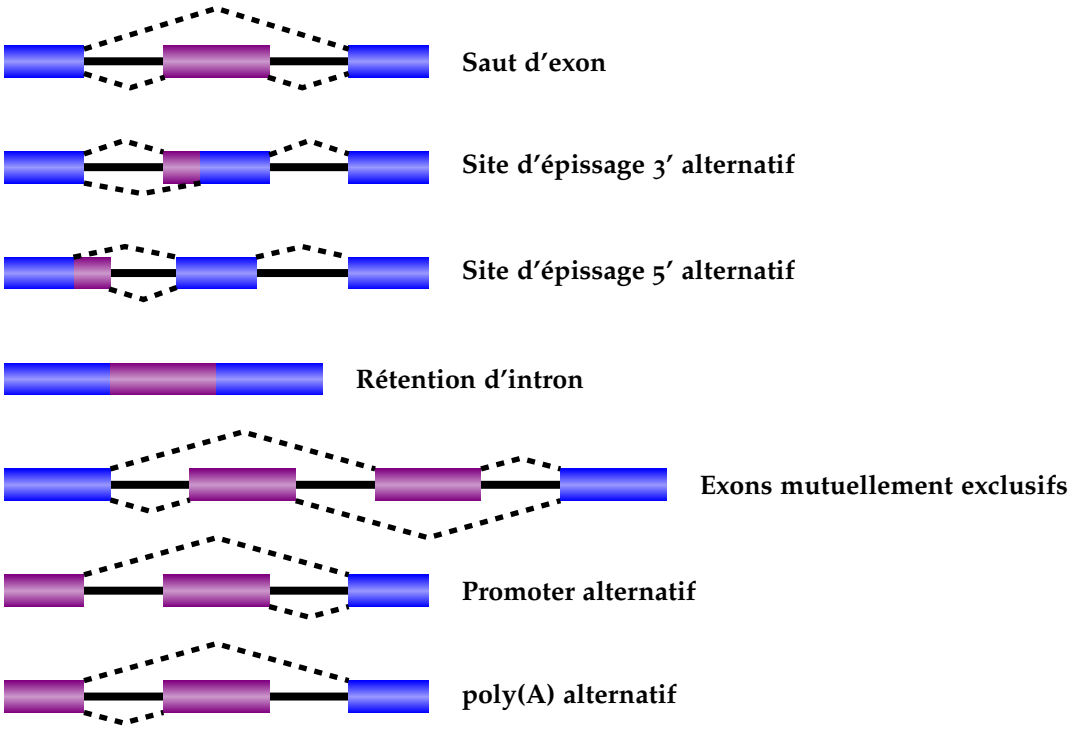


FIGURE 2.9 – *Les différents types d'épissages alternatifs (adapté de Keren et al. (2010))*

Problématiques et moyens disponibles 3

Dans ce chapitre, nous définissons les problématiques abordées dans le contexte qui vient d'être défini et présentons les moyens disponibles pour les étudier. Nous passons en revue le principe des différentes technologies développées pour le séquençage deuxième génération. Dans les années 2000, trois technologies se disputaient le marché du séquençage nouvelle génération : Roche, Illumina et Life. Nous décrivons les technologies utilisées dans ce projet dans la dernière section.

3.1 Définition des problématiques

Les résultats obtenus ces dernières années dans l'étude de la leucémie myélomonocytaire chronique, des néoplasmes myéloprolifératifs et des syndromes myélodysplasiques font émerger plusieurs questions. Ces questions sont formulées ci-dessous. Notre travail s'est efforcé de répondre à certaines d'entre elles.

1. Pourquoi des maladies du même tissu partageant des anomalies génétiques ont-elles un phénotype parfois si différent ? Des éléments de réponse ont été apportés par Itzykson et al. (2013b) mais il reste des zones d'ombre.

2. Comment de multiples combinaisons de mutations somatiques conduisent-elles à un phénotype unique ? Cela est-il dû à une altération génétique commune à tous les patients non encore identifiée ? Nous avons exploré l'ensemble des parties codantes par séquençage d'exomes entiers (WES, Whole Exome Sequencing) puis non codantes du génome par séquençage de génomes entiers (WGS, Whole Genome Sequencing) dans les monocytes des patients pour tenter de répondre à cette question.

3. Est-ce que l'analyse des cellules matures du clone permet d'identifier les mutations véritablement responsables de la maladie et de sa progression ? Il semble que les altérations s'accumulent, sans disparaître, au cours du temps. Les monocytes contiendraient toutes les anomalies génétiques accumulées au cours de la pathologie.

	HiSeq2000	454 GS FLX	SOLiDv4	Sanger 3730xl
Méthode de séquençage	Synthèse	Pyroséquençage	Ligation et codage par paires	Didésoxyribonucléotides
Longueur Read (bp)	50 SE [1], 50PE [2], 101PE	700	50+35 or 50+50	400-900
Précision (%)	98	99.9	99.94	99.999
Nombre Reads	3G	1M	1200-1400M	
Quantité de données	600 Gb	0.7 Gb	120 Gb	1.9-84 Kb
Durée	3-10j	24h	7j pour SE, 14j pour PE	20 min - 3h
Prix	$690,000	$500,000	$495,000	$95,000
Préparation de la library automatisée	Oui	Oui	Oui	Non
Coût / million bases	$0.07	$10	$0.13	$2400
Avantages	Longueur Read, rapide	Débit	Précision	Qualité, longueur des reads
Inconvénients	Assemblage des reads courts	Homopolymère≥6, coût, bas débit	Assemblage des reads courts	Coût, bas débit

TABLE 3.1 – *Comparaison des principales technologies de séquençage NGS*

4. Quelle est l'origine de l'hétérogénéité cellulaire de la leucémie myélomonocytaire chronique ? Par exemple, pourquoi une partie seulement des patients génèrent des *MDSC* ? Pourquoi les agents hypométhylants n'améliorent-ils l'hématopoïèse que chez 30-40% des patients ?

5. Les mutations n'expliquent pas tout. Les altérations épigénétiques, mais aussi le microenvironnement dans la niche médullaire et le système immunitaire jouent un rôle dans l'émergence et l'expansion du clone leucémique. Ces rôles sont très mal connus.

6. Pourquoi les agents déméthylants n'éradiquent-ils pas la maladie ? Est-ce que le traitement permet d'éliminer une partie des cellules mutées ou des sous-clones particuliers, au moins chez les répondeurs ? Ce traitement permet-il réellement de diminuer le niveau de méthylation de l'*ADN* des cellules malades et cela se traduit-il par la ré-expression de certains gènes ?

3.2 Technologies disponibles pour répondre à ces problématiques

Pour répondre à ces questions, nous avons analysé les cellules tumorales de patients à grande échelle à l'aide du séquençage nouvelle génération. Les technologies de séquençage actuelles ont été introduites sur le marché dans les années 2000 pour répondre à la limitation que présente la méthode standard Sanger. Celle-ci permet de lire des zones comprenant jusqu'à 1000 paires de bases. Le séquençage d'un génome ou même d'un exome humain ne sont ainsi pas envisageables en un temps limité. Cette nouvelle génération de séquençage est significativement plus rapide,

1. SE : Single End
2. PE : Paired End

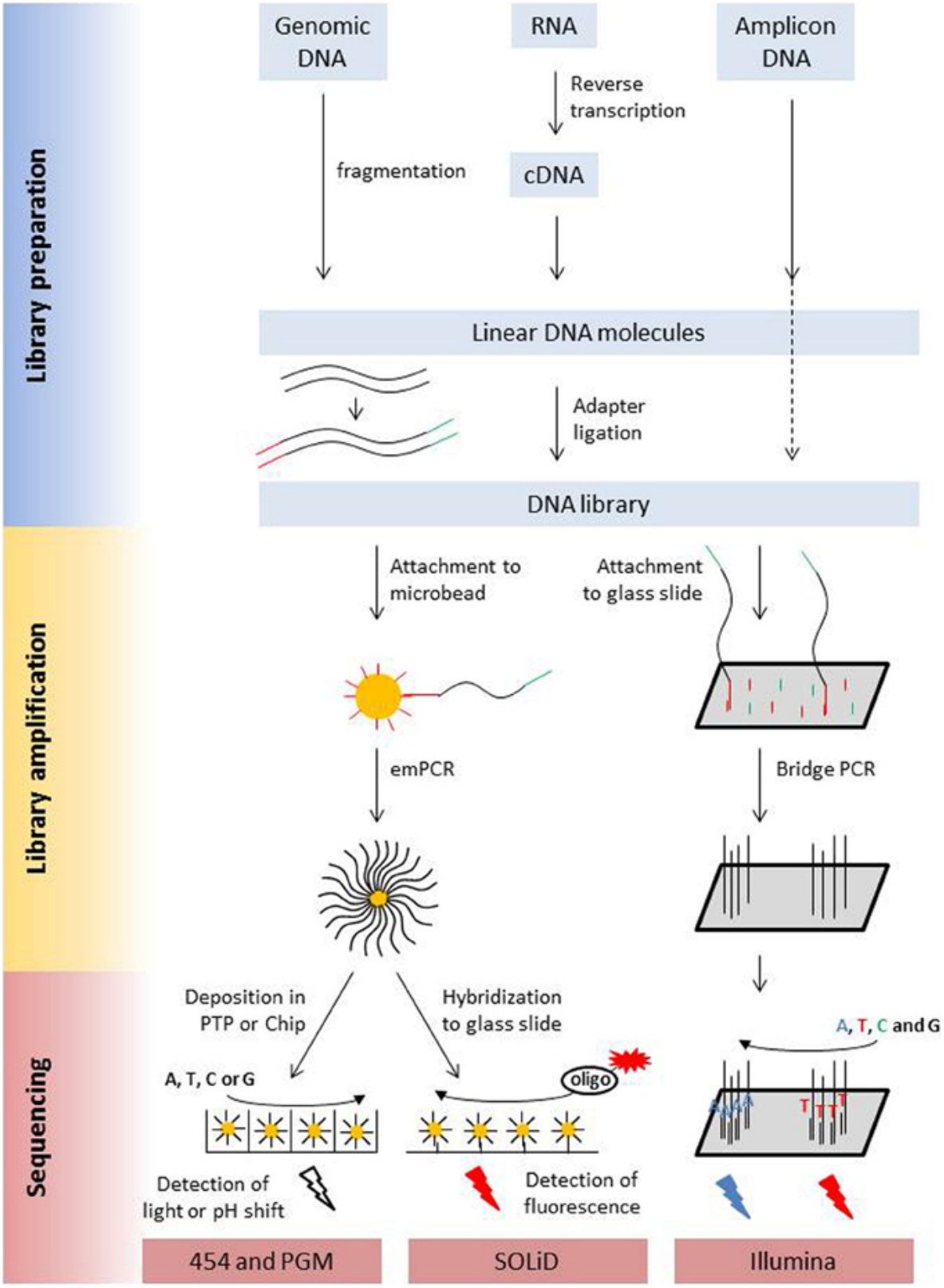


FIGURE 3.1 – *Mode de fonctionnement des principales technologies de séquençage (Knief (2014))*

moins chère, plus précise et nécessite moins de matériel (tableau 3.1). Ce séquençage est dénommé de plusieurs façons, séquençage nouvelle génération, séquençage massif, séquençage deuxième génération, séquençage à très haut débit ou séquençage *NGS*.

Au moment où ce projet a été initié, il existait trois principales technologies : 454 Lifesciences/Roche (Margulies et al. (2005)), Illumina/Solexa (Bennett (2004), Bennett et al. (2005), Bentley (2006)) et Life Technologies. Les machines disponibles étaient celles de 454 (GS FLX et GS Junior), Illumina (GA IIx et HiSeq2000) et Life (SOLiD)[3]. La figure 3.1 présente le mode de fonctionnement et le tableau 3.1 fournit les différentes caractéristiques de ces machines. Ces différentes techniques reposent sur trois étapes : préparation des librairies d'ADN génomique (ADNg) ou ADN complémentaire (ADNc), amplification des molécules par *PCR* et réaction de séquençage. Elles reposent sur des principes différents et sont sujettes à des types d'erreurs spécifiques. Une description et comparaison des trois méthodes est fournie par Morozova and Marra (2008), Mardis (2008), Metzker (2009) et une comparaison entre Illumina et Roche par Luo et al. (2012). Ces deux dernières années, les technologies Life puis Illumina se sont développées autour du séquençage ciblé, permettant le séquençage de régions courtes (de quelques dizaines à quelques centaines de bases) de nombreux échantillons simultanément.

Technologie Illumina

Nous commençons par décrire la technologie Illumina/Solexa car c'est essentiellement celle-ci qui a été utilisée dans le travail rapporté ici. La société Illumina détient environ 60% du marché, malgré plusieurs inconvénients bien connus : lectures courtes, fréquentes évolutions techniques et logicielles, taille des fichiers générés. Illumina commercialise plusieurs machines, GAIIx (arrêté en 2014), HiSeq2000, et plus récemment HiSeq2500, HiSeq3000 et HiSeq4000 ainsi que des machines de plus petites capacités, comme le MiSeq, et de plus grandes capacités par mise en réseau de plusieurs séquenceurs, comme le HiSeqX Ten (table 3.2).

Quelle que soit la machine, il s'agit d'un séquençage par terminaison cyclique réversible. Voici plus en détails une description des étapes (figure 3.2) :

1. Construction des libraries : les molécules d'*ADN* sont fragmentées par sonication, par des enzymes ou par utilisation d'un gaz. Les fragments sont réparés afin que leurs extrémités soient franches, puis un nucléotide A est ajouté pour pouvoir lier des adaptateurs. Les fragments sont sélectionnés par leur taille moléculaire, de manière à avoir des fragments de même longueur.
2. Amplification : les fragments sont déposés sur une plaque appelée flowcell, où sont accrochés des adaptateurs, complémentaires à ceux de la library. Le matériel disposé sur la plaque varie en fonction du séquençage que l'on souhaite réaliser. L'amplification par ponts peut commencer. L'*ADN* est alors dénaturé. Des clusters sont formés, comprenant chacun une même séquence d'*ADN* et sa séquence complémentaire des millions de fois. Pour ne garder que la molécule d'intérêt, une enzyme vient découper la séquence complémentaire. Le séquençage peut commencer.
3. Séquençage : une *ADN* polymérase ajoute un nucléotide modifié porteur d'une molécule fluorescente et d'un terminateur labile. Chaque nucléotide porte un fluorochrome particulier. Suite à l'incorporation, le terminateur empêche l'ajout d'une nouvelle base. Le milieu réactionnel est lavé pour supprimer les nucléotides non incorporés. Les fluorochromes sont ensuite excités. Les émissions lumineuses sont détectées par le séquenceur puis traduites en une séquence. A la fin de chaque cycle, le fluorochrome et le terminateur en 3' sont clivés

3. Sequencing by Oligonucleotide Ligation and Detection

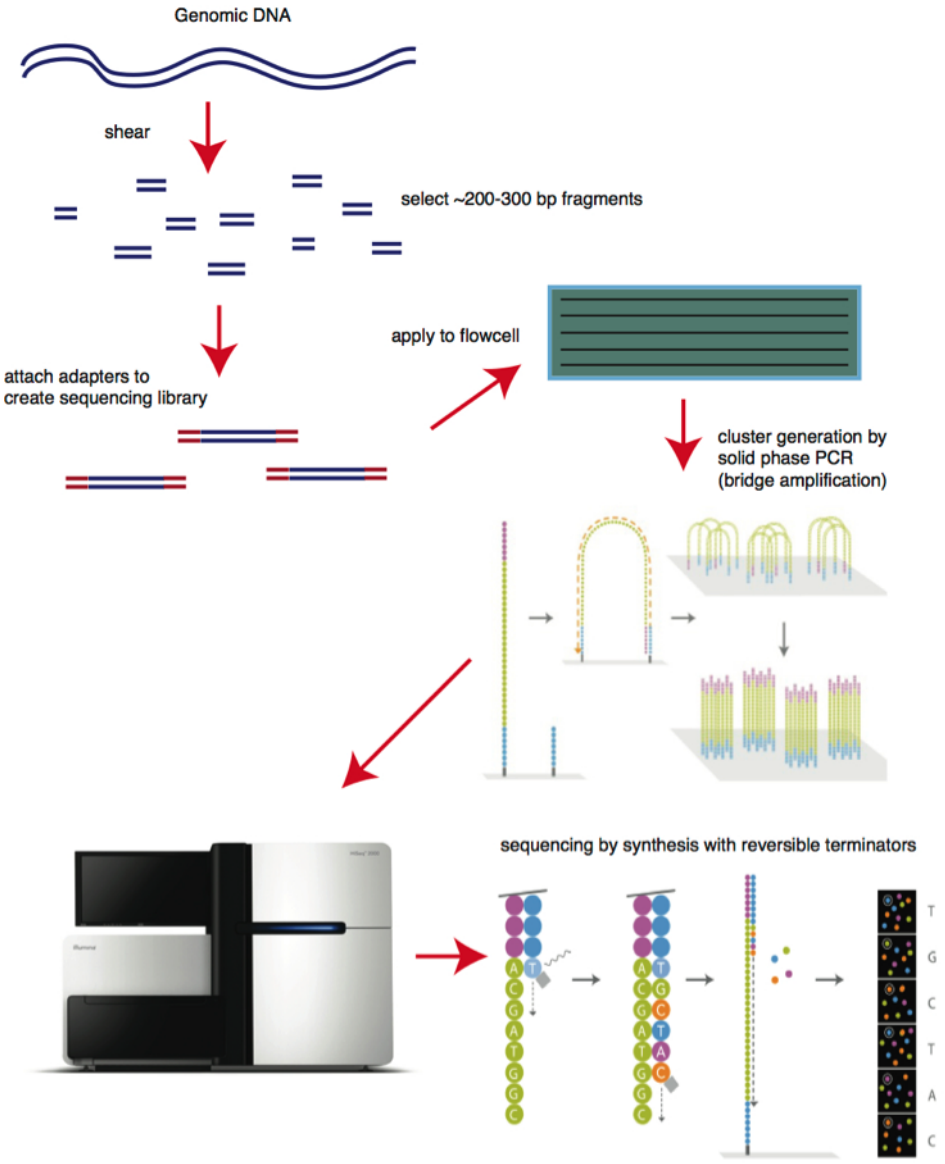


FIGURE 3.2 – *Description du principe de la technologie Illumina (http ://bitesizebio.com/13546/sequencing-by-synthesis-explaining-the-illumina-sequencing-technology)*

pour que la réaction de séquençage puisse continuer.

Cette technologie présente deux biais principaux. La taille des lectures générées est plutôt faible. Ceci tend toutefois à s'améliorer : les lectures que nous avons générées font 101bp, les séquenceurs HiSeq3000 et HiSeq4000 permettent de lire des lectures de 150 nucléotides, ce qui reste moins que les 400 nucléotides lus par la méthode de pyroséquençage. Le deuxième biais est la diminution de la fiabilité du séquençage aux extrémités des lectures. Plus le nombre de cycles augmente, plus les décalages dans la séquence s'accumulent conduisant à une augmentation du bruit de fluorescence et une interprétation erronée des signaux lumineux.

Technologie Life

Life technologies a développé plusieurs séquenceurs (décrits brièvement table 3.3), la machine SOLiD pour du séquençage classique n'est plus utilisée, et deux séquenceurs dédiés au séquençage ciblé.

Comme pour les autres technologies, Life repose sur trois principales étapes et se rapproche le plus de la technologie 454 décrite ci-après. Nous nous limitons ici à la description du fonctionnement de l'Ion Torrent PGM.

1. Construction des libraries : des *PCR* des régions cibles sont réalisées. Des adaptateurs sont ajoutés aux produits des *PCR* et chaque molécule est placée sur une microbille magnétique d'1 micron d'épaisseur *via* un adaptateur sur cette dernière.
2. Amplification : les molécules sont amplifiées sur les microbilles par *PCR* en émulsion. Chaque bille est placée dans un seul puits sur une lame de verre, et non dans des microcuves comme dans la technologie 454.
3. Séquençage : détection du pH par le capteur dans chaque cavité et rinçage par microfluidique pour chaque base. Chaque cavité permet de séquencer une molécule d'environ 200 bases.

Le séquençage ciblé nécessite une étape préalable pour la sélection des régions cibles : construction d'amorces spécifiques à des régions cibles du génome. La figure 3.3 compare les deux méthodes de séquençage ciblé utilisées dans ce projet.

Technologie 454 Life sciences

La technologie 454 a été la première sur le marché et repose comme les autres sur trois étapes :

1. Construction des librairies : l'*ADN* de l'échantillon est fractionné aléatoirement. Les fragments d'*ADN* double brin sont réparés pour posséder des extrémités compatibles avec les extrémités des adaptateurs. Les adaptateurs nécessaires à la *PCR* en émulsion et à la réaction de séquençage (adaptateur A et adaptateur B) peuvent ainsi être ajoutés à l'*ADN*. Les fragments peuvent être bornés par deux adaptateurs A, deux adaptateurs B ou un de chaque. Seuls les fragments possédant un adaptateur A et un adaptateur B pourront être amplifiés lors de la *PCR*. Ce seront les seuls séquencés. Les fragments sont ensuite mis en contact avec des billes. Les fragments bornés par deux adaptateurs A ne pourront pas être récupérés sur ces billes. L'*ADN* fixé sur ces billes subit alors une dénaturation et seules les molécules d'*ADN* simple brin bornées par un adaptateur A et un adaptateur B pourront se décrocher des billes. Les fragments possédant deux adaptateurs B restent piégés.
2. Amplification : l'*ADN* simple brin est mis en contact avec des billes sur lesquelles sont fixées par l'extrémité 5' des amorces A (billes A) ou B (billes B) respectivement complémentaires de la séquence de l'un ou l'autre des adaptateurs. Un seul fragment est fixé sur chaque bille. Les

	MiSeq	NextSeq 500		HiSeq				HiSeq X series	
Versions	1/2/3			2000	2500		3000	4000	X Five	X Ten
Mode		Débit moyen	Haut débit		Run rapide	Haut débit				
Flowcell(s) par run	1	1	1	1 ou 2	1 ou 2	1 ou 2	1	1 ou 2	1 ou 2	1 ou 2
Quantité de données (Gb)	0.3-15	20-39	30-120	100-600	10-300	50-1000	125-750	125-1500	900-1800	900-1800
Durée	5-55h	15-26h	12-30h	2-11j	7-60h	1-6j	1-3.5j	1-3.5j	3j	3j
Reads par Flowcell (millions)	25	130	400	3000	300	2000	2500	2500	3000	3000
Max longueur read (bp)	2*300	2*150	2*150	2*100	2*250	2*125	2*150	2*150	2*150	2*150

TABLE 3.2 – *Caractéristiques des différents séquenceurs Illumina*

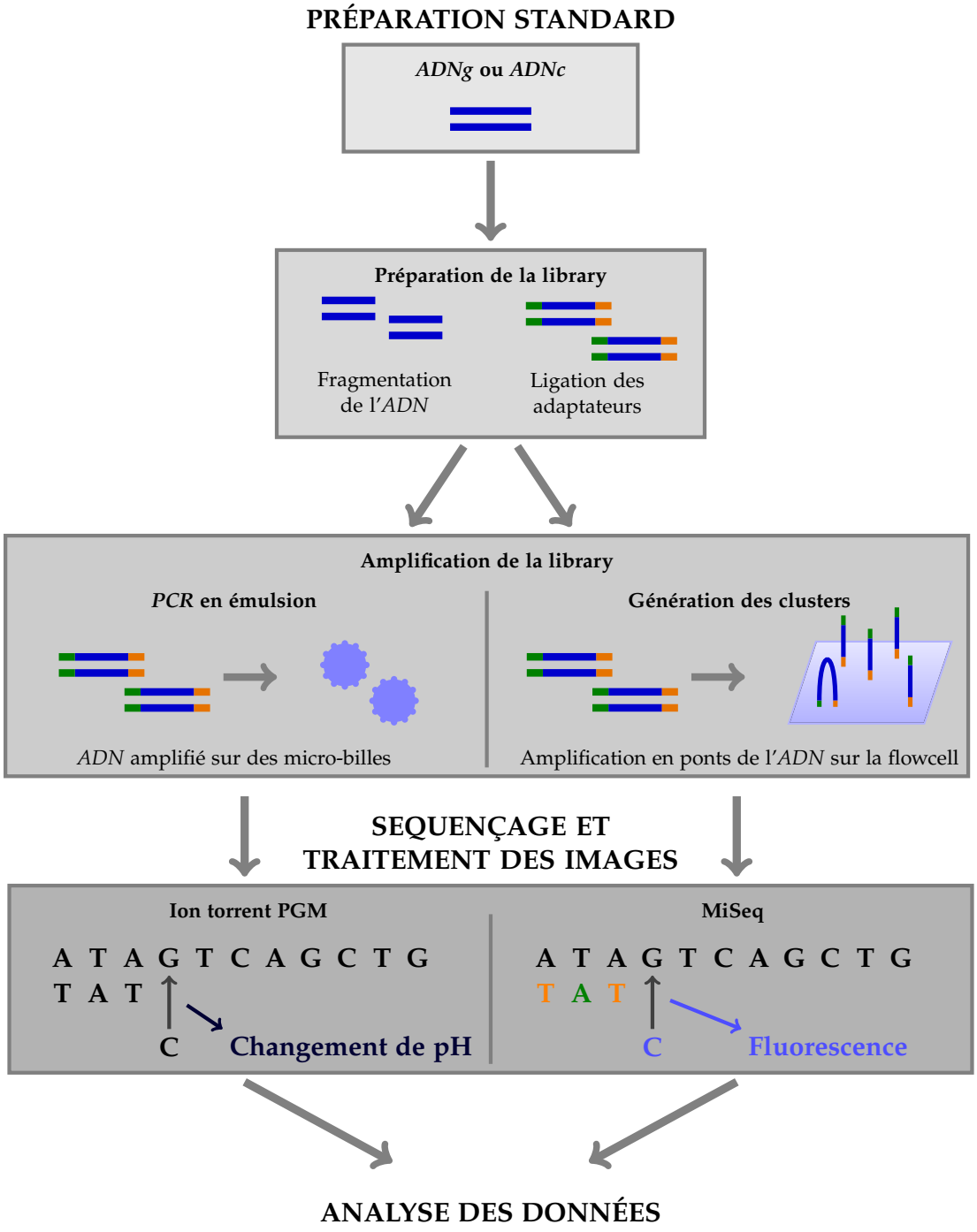


FIGURE 3.3 – *Principe des méthodes de reséquençage (adapté de Grada and Weinbrecht [59])*

Séquenceur	Ion Torrent	Ion Proton	Sanger	SOLiD
Description	Précis, simple et rapide pour le séquençage ciblé ou microbien	Haute qualité, séquençage *WES* et *RNASeq*	Gold standard pour les analyses rapides	Pour des projets de petite à grande envergure
Données générées	de 30Mb à 2Gb	$\leq$14Gb		
Vitesse	2h	2h	1 à 96 échantillons en $\leq$5h	
Reads (bp)			$\leq$1000	
Application	*ARN* ciblé *ADN* ciblé *ADN* microbien	Transcriptome *WES* *WGS* *RNASeq*	Confirmation de résultats Reséquençage ciblé Identification de bactéries, de champignons et de virus Microsatellite	*WGS* *WES* *RNASeq*

TABLE 3.3 – *Caractéristiques des différents séquenceurs Life*

billes sont ensuite mises dans une émulsion contenant les composés nécessaires à la réaction de polymérisation en chaîne (PCR en émulsion) et permettant d'individualiser chaque bille. Chaque goutte d'émulsion devient un "microréacteur" dans lequel l'unique fragment d'*ADN* fixé sur la bille va être amplifié. Au cours du premier cycle de *PCR*, l'amorce associée au fragment d'*ADN* initial assure la synthèse de son brin complémentaire. Au cycle suivant, après dénaturation, la molécule d'*ADN* simple reste fixée sur la bille et est utilisée pour la synthèse d'une autre matrice à partir de l'autre amorce présente dans le milieu réactionnel. La molécule d'*ADN* simple brin de départ est réutilisée par une autre amorce fixée sur la bille.

3. Séquençage : après amplification, les billes sont déposées sur une plaque contenant plusieurs millions de puits dont les diamètres permettent de ne récupérer qu'une seule bille. La réaction de pyroséquençage se réalise au sein de chacun de ces puits. Les nucléotides sont rajoutés de manière séquentielle. Après chaque ajout d'un nucléotide, un traitement permet d'éliminer le surplus, puis le nucléotide suivant est incorporé et ainsi de suite. Lorsqu'un nouveau nucléotide est ajouté, l'ATP sulfurylase utilise le pyrophosphate relâché lors de la polymérisation pour générer de l'ATP. Cet ATP est utilisé par la luciférase pour oxyder la luciférine en oxyluciférine et émettre de la lumière. C'est ce signal lumineux qui est détecté par une caméra puis traduit en chromatogramme. Cette méthode permet de séquencer des reads de 400bp.

Si un nucléotide est incorporé plusieurs fois dans le même cycle, un signal lumineux proportionnel au nombre de nucléotides est décelé. Le principal défaut de cette méthode est qu'au-delà de 7 nucléotides environ, le signal est difficilement exploitable.

Chaque nucléotide détecté par l'une de ces méthodes se voit attribuer un score de qualité. Ce score va généralement de 4 à 60 avec les valeurs les plus grandes correspondant aux plus grandes

FIGURE 3.4 – *Séquenceur MinION : l'avenir du diagnostic moléculaire ? (http ://www.traqueur-stellaire.net/2014/09/sequenceur-adn-cle-usb)*

qualités. Il représente la probabilité que la base détectée soit fausse. Lorsque l'on connaît le score de qualité Q, la probabilité d'erreur P s'obtient par : $P = 10^{\frac{-Q}{10}}$.

Plusieurs sociétés de séquençage se développent, comme Pacific Biosciences et Oxford Nanopore, autour du séquençage troisième génération.

Pacific Biosciences développe depuis plusieurs années des séquenceurs (PacBio RS et PacBio RS II) permettant de générer des lectures de quelques milliers de bases. Ces séquenceurs ne sont toujours pas officiellement sur le marché mais sont utilisés pour des tests dans quelques centres. Ils sont utilisés, en combinaison avec des séquenceurs Illumina, pour faire de l'assemblage *de novo* de génomes inconnus. Leur principal inconvénient est un taux d'erreur important, de l'ordre de 15%.

Oxford Nanopore a mis en test en 2015 le MinION, un séquenceur de la taille d'une clé USB, permettant une connexion directe sur PC *via* un port USB. L'échantillon à séquencer est directement déposé au sein du séquenceur (figure 3.4). MinION permet de séquencer de courts brins d'*ADN* en quelques secondes aussi bien que des brins de quelques milliers de bases en quelques heures. Cette technologie serait idéale pour une utilisation en clinique, car elle permettrait d'obtenir le statut moléculaire de quelques cibles en 1 journée. Pour le moment, le taux d'erreur serait important. Affaire à suivre...

3.3 Technologies utilisées

Le séquençage des échantillons d'*ADN* (exomes, génomes) et d'*ARN* a été fait à l'aide d'un séquenceur Illumina HiSeq2000. Il faut souligner que le séquenceur HiSeq2000 a un temps de séquençage de 10 jours environ pour des lectures paired-ends de 100bp alors que le HiSeq4000 réalise le séquençage paired-ends de 150 bp en 3 jours. La validation des mutations détectées dans les échantillons d'exomes a été réalisée avec un séquenceur Ion Torrent PGM. Les principaux désavantages de ce séquenceur sont la difficulté à séquencer les homopolymères et le coût. Aussi avons-nous par la suite abandonné le PGM au profit du séquenceur Illumina MiSEQ. Les gènes candidats ont été séquencés avec le séquenceur MiSeq.

Pour séquencer des exomes, des kits de capture sont utilisés. Ces kits évoluent très rapidement dans le but de capter l'ensemble des régions codantes à une profondeur uniforme. Deux types de kits existent : les kits SureSelect (Agilent) et les kits TruSeq (Illumina). Dans notre projet où nous avons séquencé des échantillons trois années durant, les kits ont évolué. Ainsi, nous avons testé

les kits TruSeq sur 3 patients. La capture ne donnant pas une couverture suffisante, l'exome de tous les autres échantillons a été capturé par des kits SureSelect. Les exomes de 3 patients ont été capturés par des kits de version 2 ; 11 par la version 3 ; 2 par la version 4 comprenant les régions non traduites 5' et 3'. Cette version a généré beaucoup de variants *a priori* artefactuels dans les régions non codantes. L'exome de la majorité des patients a été capturé par la version 4 (19) ou la version 5 (11). Une nouvelle version est disponible depuis peu : la version "5 clinic", qui est supposée assurer une profondeur minimum dans les régions codantes, de manière à pouvoir être utilisé à des fins diagnostic.

Pour le reséquençage, il n'est plus question de capture. Des amorces sont synthétisées de manière à créer par *PCR* des amplicons correspondant aux régions que l'on souhaite séquencer. L'*ADN* est mis en présence de ces amplicons pour sélectionner les régions cibles puis est amplifié par *PCR*.

Pour le génome, la procédure est plus simple : l'*ADN* est fragmenté et les librairies sont faites avec la totalité du matériel. Il n'y a pas d'étape de sélection de régions particulières.

Pour capter les anomalies transcriptomiques, l'*ADNc* est séquencé car l'*ARN* en lui-même est trop fragile et le nombre important de cassures créerait des séquences chimères. L'*ARN* est composé principalement d'*ARNm* ($\leq$5% de l'*ARN* total), d'*ARN* de transfert (5%-10% de l'*ARN* total) et d'*ARN* ribosomique (environ 80% de l'*ARN* total). Suivant la question posée, certains *ARN* doivent être analysés. De ce fait, il existe plusieurs protocoles de capture de l'*ADNc*. Le plus simple, qui est le moins utilisé, est d'explorer l'ensemble des *ARN*, ce qui ne nécessite pas d'étape de sélection positive ou déplétion. Les deux protocoles les plus utilisés sont la déplétion en *ARN* ribosomaux et la sélection des *ARN* possédant une queue polyA, *i.e.* la majorité des *ARNm* ainsi que certains *ARN* non codants. Dans le cas de la déplétion en *ARN* ribosomaux, les *ARN* ribosomaux sont extraits de l'*ADNc* total et le reliquat (environ 20% de l'*ARN* total) est utilisé pour la suite du protocole, afin d'analyser les *ARN* restants. Si le but est l'étude des *ARNm*, les *ARN* possédant une queue polyA sont extraits (environ 5% de l'ARN total) et utilisés pour la suite.

Notre expérience de séquençage d'ARN (RNA-Seq, RNA Sequencing) visant à étudier les dérégulations entre patients *LMMC* et sujets sains a été réalisée en 2013 en utilisant l'*ADNc* déplété en ribosomes, seul protocole alors appliqué dans la plateforme de génomique de Gustave Roussy. Étant donné que la sélection des *ARN* possédant une queue polyA permet de se concentrer sur les *ARN* qui vont être traduits, ce protocole est aujourd'hui le plus utilisé car pour une même couverture, les *ARN* d'intérêt sont plus couverts. L'expérience visant à étudier l'effet du traitement chez des patients non répondeurs et chez des patients stables a été réalisée fin 2014 sur des *ARN* avec queue polyA, le protocole ayant été mis au point entre temps à la plateforme de génomique.

Deuxième partie

Méthode

Le séquençage à très haut-débit de génomes et en particulier d'exomes est aujourd'hui répandu dans de nombreux domaines, en génétique, en cancérologie, en phylogénétique par exemple, en raison de la diminution importante des coûts de séquençage. Cette technique rencontre un véritable engouement depuis 2009. L'implication de mutations du gène *DHODH* dans le syndrome de Miller (Ng et al. (2009)) est l'une des premières découvertes rendues possibles grâce au séquençage exomique dans les maladies mendéliennes rares. Cette technique a également permis, dès ses débuts, de mettre en évidence la mutation du gène *IDH1* dans les leucémies aiguës myéloïdes (Mardis et al. (2009)) et de considérer ce gène comme un suppresseur de tumeur, inactivé par des mutations dominantes de *R132*.

Cette technologie permet le séquençage *de novo* ou le re-séquençage d'un génome connu, l'annotation (ou la ré-annotation) de plus en plus précise d'un génome, l'étude de la variabilité génétique entre individus d'une même espèce et la détection de mutations à l'origine de diverses pathologies. L'étude du transcriptome tire également profit de cette nouvelle technologie : identification des sites de démarrage de la transcription, des séquences frontières intron/exon, étude de l'épissage alternatif, analyse du niveau d'expression des gènes ou encore étude des petits ou longs *ARN* non codants par exemple. Enfin, l'étude du profil de méthylation, des interactions ADN-protéines et des modifications post-traductionnelles des histones sont également possibles par ce procédé.

La nouvelle génération de séquençage à très haut-débit regroupe l'ensemble des technologies développées depuis 2005. Actuellement, nous sommes entre les technologies à très haut-débit dites de deuxième génération, qui nécessitent une étape d'amplification avant le séquençage, et les technologies à très haut-débit dites de troisième génération, ne nécessitant plus l'amplification des molécules séquencées. A l'heure actuelle, les efforts portent sur le séquençage à l'échelle unicellulaire.

Une analyse précédemment réalisée au laboratoire a étudié 18 des gènes fréquemment mutés dans la *LMMC* chez 312 patients. Nous avons décidé d'étendre l'analyse génomique de la *LMMC* à l'ensemble des gènes afin de caractériser au mieux cette pathologie. Nous avons opté pour un séquençage de l'exome et du génome de cellules leucémiques et de cellules contrôles de patients pour rechercher les mutations somatiques, pertes d'hétérozygotie et altérations du nombre de copies. Étant donné d'une part, les travaux précédemment réalisés dans la *LMMC* et d'autre part, l'émergence des technologies à très haut-débit, nous avons initié ce projet de séquençage à grande échelle en 2011.

Face à l'émergence des données générées par les technologies à très haut-débit dans de nombreux domaines de la biologie, les méthodes permettant une analyse précise et rapide se développent depuis plusieurs années. Alors que certaines problématiques semblent avoir des solutions convenables suite au développement de plusieurs dizaines d'outils comme l'alignement de séquences d'*ADNg*, d'autres restent en plein développement, comme la quantification du niveau d'expression d'isoformes. Dans le travail mené ici, nous avons utilisé certains de ces outils pour répondre à nos problématiques.

Analyse de données de séquençage à très haut débit

4

Dans cette partie, nous décrivons en détails les analyses de données réalisées suite aux séquençage *WES*, *WGS*, ciblé et *RNASeq*.

4.1 Analyse de séquences d'ADN

L'étude de l'*ADN* a pour principale vocation la recherche de mutations et variations structurales. Les étapes nécessaires à la conversion des données brutes issues du séquençage en une liste de variants d'une seule base (SNV, Single Nucleotide Variation) et insertions/délétions (INDEL) sont schématisées figure 4.1. Nous citons, de manière non exhaustive, quelques-uns des outils utilisables. Chacune des étapes contribue à la précision du résultat final. Le processus débute par un contrôle qualité. L'une des principales étapes est l'alignement. Cela consiste à rechercher l'origine de chaque lecture générée dans un génome de référence. Pour limiter les artefacts détectés lors de la détection des variants, il est nécessaire d'effectuer la suppression des duplicats de *PCR* ainsi que le réalignement des lectures autour des *INDEL* candidates en aval de l'alignement. La recalibration des bases permet de réévaluer la qualité des bases variantes à l'aide d'une liste de polymorphismes (SNP, Single Nucleotide Polymorphism) avec plus de précision que la qualité

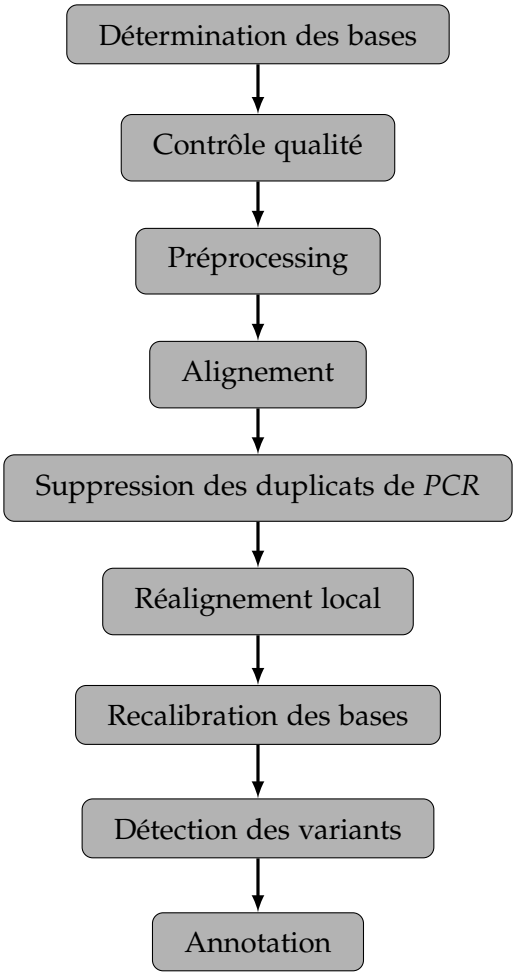


FIGURE 4.1 – *Analyse standard de séquences d'ADN pour la détection de variants*

définie par les séquenceurs. Les données sont alors prêtes pour l'étape de détection. Nous présentons différents algorithmes de détections de *SNV* et/ou *INDEL*, en détaillant en particulier les outils dédiés aux données de cancer. L'analyse de données *NGS* a été discutée dans de nombreux papiers par Dalca and Brudno (2010), DePristo et al. (2011), Nielsen et al. (2011), Pattnaik et al. (2012) ou You et al. (2012) pour n'en citer que quelques-uns. Les spécificités des données de cancer ont été abordées dans quelques articles, par Ding et al. (2010) par exemple. Les variants détectés sont ensuite annotés à l'aide d'outils d'annotation. Les mutations ainsi annotées constituent les éléments de base dans la caractérisation génomique d'un patient ou d'une pathologie. Il est alors possible de rechercher les gènes drivers ou les pathways altérés par exemple.

La recherche de mutations dans les séquences d'*ADN* est relativement similaire en ce qui concerne le traitement des données *WES*, *WGS* ou ciblées. Seule l'étape de suppression des duplicats de *PCR* varie suivant le séquençage réalisé : elle ne se fait pas dans le cas de séquençage ciblé. La figure A.1 (p169) présente en détails l'analyse réalisée sur les données *WES*, allant des données de séquençage brutes à l'obtention des variants détectés, *SNP* et *INDEL*. De même, les figures A.2 et A.3 (p170 et p171) présentent en détails l'analyse réalisée sur les données de reséquençage et sur les données *WGS* respectivement. Ces procédures emploient des outils libres et communément utilisés.

4.1.1 Contrôle qualité et préprocessing

A la fin du processus de séquençage, les données sont stockées dans des fichiers FASTQ. Ils contiennent les lectures ainsi que la qualité des bases de chaque read. Pour se faire une idée de la qualité des données générées, on procède à un contrôle qualité à l'aide de FastQC [1] par exemple. On représente classiquement la qualité de chaque base pour l'ensemble des lectures. On observe généralement une diminution de la qualité en fin de read, surtout pour les lectures en sens indirect. Si des problèmes de qualité sont observés, il faut éliminer les reads ou parties des reads de qualité insuffisante. Trimmomatic (Bolger et al. (2014)) ou cutadapt (Martin (2011)) par exemple permettent de filtrer les lectures suivant plusieurs critères qualité. Des exemples de profils qualité de données de séquençage ciblé et de génome sont donnés en annexe A.2 p172 avant et après préprocessing avec Trimmomatic.

4.1.2 Alignement de séquences d'ADN sur un génome de référence

L'alignement des lectures sur un génome de référence consiste à rechercher à quelle position du génome correspond chaque lecture générée. L'alignement est fondamental car toutes les autres étapes dépendent de son bon déroulement. De nombreux outils ont été développés pour répondre à cette problématique : MAQ [2] (Li et al. (2008a)), SOAP [3] (Li et al. (2008b)), BWA [4] (Li and Durbin (2009)), Bowtie (Langmead et al. (2009)), SOAP2 (Li et al. (2009c)), Novoalign [5], SHRIMP [6] (Rumble et al. (2009)), SHRIMP2 (David et al. (2011)), Bowtie2 (Langmead and Salzberg (2012)), SOAP3 (Liu et al. (2012)), ... Le choix de l'outil doit être guidé par le type de données, la technologie de séquençage utilisée, le but escompté, la taille de reads (petite, grande, uniforme, variable...), la gestion des alignements multiples et du mode paired-end, la gestion ou non des *INDEL*, l'utilisation par la communauté et surtout la maintenance de l'outil.

L'alignement se fait en deux étapes. Dans un premier temps, soit la séquence de référence, soit l'ensemble des lectures séquencées est indexé afin de permettre une recherche rapide. La majorité des aligneurs actuels indexe le génome de référence plutôt que les lectures séquencées, parce que c'est plus économe en mémoire vive et en temps de calcul. Dans un deuxième temps, l'alignement en lui-même se fait en recherchant les lectures dans l'index créé.

La diversité des outils existant s'explique par les multiples solutions proposées au problème d'indexation des données. Étant donné que l'on cherche la provenance de millions de lectures (typiquement de 50 à 150 bp) dans une référence qui contient environ 3.1 milliards de bases chez l'Homme, il est indispensable d'optimiser cette recherche. Le génome de référence est transformé, on parle d'indexation, de manière à nécessiter peu d'espace de stockage et à pouvoir être interrogé de manière extrêmement rapide. Il existe deux principales méthodes d'indexage : le "suffix index" et les tables de hachage. Les aligneurs basés sur le principe du "suffix index" sont économes en mémoire, plus rapides mais moins précis que les algorithmes basés sur les tables de hachage.

Nous avons utilisé une méthode d'alignement basée sur le principe du "suffix index". Cette méthode se décline en quatre approches : "le suffix try", le "suffix tree", le "suffix array" et l'index FM (figure 4.2). Les deux méthodes basées sur des arbres sont similaires. Le "suffix tree" est optimisé par rapport au "le suffix try" afin de nécessiter moins d'espace pour stocker la référence. Néanmoins, cette méthode nécessiterait environ 47Go pour stocker le génome humain en mé-

1. http ://www.bioinformatics.babraham.ac.uk/projects/fastqc/
2. Mapping and Assembly with Qualities
3. Short Oligonucleotide Analysis Package
4. Burrows-Wheeler Alignment
5. www.novocraft.com
6. SHort Read Mapping Package

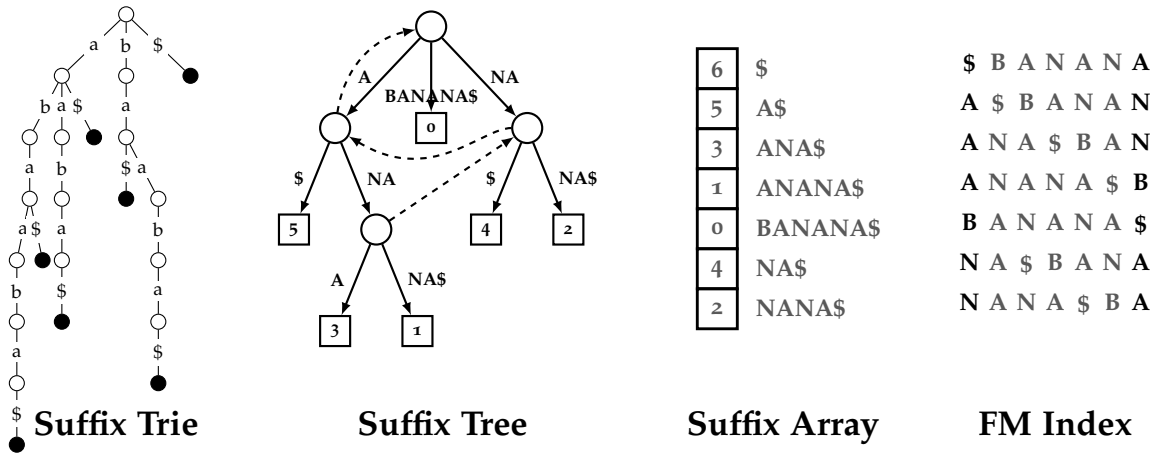


FIGURE 4.2 – *Les quatre méthodes basées sur le "suffix index" (exemple de Ben Langmead)*

moire, ce qui n'est pas envisageable. Les méthodes "suffix array" et index FM sont différentes des approches basées sur les arbres. Elles permettent une nette diminution des ressources nécessaires pour stocker la référence et utiliser l'index. Le "suffix array" permet de stocker le génome de référence avec 12Go de RAM et l'index FM permet de faire nettement mieux, avec seulement 1.5Go de RAM. C'est pourquoi les outils les plus récents utilisant le principe du "suffix index" ont recours à l'index FM, comme c'est le cas de BWA et Bowtie. Pour la description de ces méthodes, de nombreuses publications sont consultables et des vidéos de Ben Langmead, le développeur de Bowtie et Bowtie2, sont disponibles en ligne.

L'alignement génère un fichier SAM ((Sequence Alignment/Map) format), qui contient, en plus des informations contenues dans les FASTQ, les positions auxquelles les reads sont alignés et leur score de qualité de l'alignement. Notons que pour les organismes n'ayant pas de génome de référence, l'alignement est impossible. Dans ce cas, les reads sont assemblés en s'aidant éventuellement du génome d'une espèce voisine ; il s'agit d'une construction *de novo*. L'assemblage consiste à grouper plusieurs lectures, créant ainsi de plus grands ensembles de nucléotides appelés contigs, qui sont assemblés au fur et à mesure jusqu'à obtenir un unique contig, à savoir le génome étudié. La détection de longues *INDEL* est ainsi possible et en général les *INDEL* sont plus justement décelées que lors d'un alignement sur un génome de référence. Notons qu'il est possible de réaliser un assemblage suivi d'un alignement dans le but par exemple d'améliorer la détection d'*INDEL* même quand un génome de référence est disponible. Mais l'assemblage n'est que peu souvent réalisé lorsqu'un génome de référence existe en raison des temps de calcul et surtout de la mémoire vive nécessaire car plusieurs centaines de Go sont nécessaires.

Le génome de référence utilisé pour l'alignement a été hg19 (NCBI human genome assembly build 37). L'alignement a été réalisé avec BWA-backtrack version 0.5.9 pour les exomes et BWA-MEM version 0.7.5a pour les génomes. Cette différence s'explique par le fait que les données *WGS* étant bien plus volumineuses que les données *WES*, l'alignement était long (4 jours dans nos conditions d'utilisation). L'utilisation de l'algorithme MEM, qui s'était répandu entre le moment où nous avons initié l'analyse des données de *WES* et celui où nous avons analysé les données de *WGS*, a permis de diviser ce temps par 4. Une fois l'alignement réalisé, nous avons vérifié la qualité de l'alignement. Nous avons calculé le pourcentage de lectures qui s'alignent de manière unique sur la référence, qui sont les lectures utilisées pour la suite, et les pourcentages de lectures qui s'alignent de manière non unique ou qui ne s'alignent pas sur la référence. Nous avons calculé la couverture moyenne de chacun des échantillons et pour avoir une idée plus précise du pourcentage de régions suffisamment couvertes, nous avons calculé le pourcentage de bases couvertes

par 1, 10, 20, 50 et 100 lectures. Avec ces informations, nous pouvons dire si un échantillon est couvert suffisamment ou non et donc si on risque ou non de manquer de nombreuses mutations.

4.1.3 Suppression des duplicats, réalignement et recalibration

Les fichiers SAM obtenus à l'étape d'alignement ont été compressés en fichiers BAM avec l'outil Picard, versions 1.76 et 1.112 pour les données d'exome et génome respectivement. Les fichiers BAM ont été triés et indexés avec Picard. Les lectures parfaitement identiques s'alignant exactement à la même position dans le même sens de lecture, appelées duplicats de *PCR*, proviennent de l'amplification par *PCR* et non d'un fractionnement identique d'une partie de l'*ADN*. Il convient de les supprimer puisque s'ils contiennent une erreur de séquençage, cette erreur voit sa fréquence augmenter si bien qu'elle peut être confondue avec un variant. Les duplicats ont été supprimés avec Picard. L'intérêt de ces étapes est discuté par Jia et al. (2012). SAMtools permet également la réalisation de ces étapes. Nous ne l'avons pas utilisé car certains défauts ont été rapportés dans la gestion de paires de lectures.

Les problèmes d'alignement se produisent particulièrement dans les régions répétées et autour des *INDEL* et induisent le mismatch de plusieurs bases près du site de mésalignement. Ces mismatchs peuvent être détectés à tort comme des *SNP* lors de la détection. Nous avons donc procédé à un ré-alignement des lectures autour des *INDEL* avec GATK [7] version 2.0-39 et 3.1 pour le *WES* et le *WGS* respectivement. Dans un premier temps, nous avons recherché les régions dans lesquelles réaliser un alignement plus précis en ciblant les *INDEL* répertoriées dans les bases de données publiques ou les *INDEL* détectées dans les échantillons analysés. Puis, nous avons effectué le réalignement en lui-même dans ces régions. Nous avons finalement recalibré les bases, de manière à corriger les biais des scores de qualité des séquenceurs. Pour cela, nous avons fourni une liste de *SNP* connus chez l'Homme à GATK. Il recherche les nucléotides différents de la référence. Si les variants ne sont pas répertoriés, ils sont considérés alors comme des erreurs de séquençage et sont réévalués en prenant en compte leur contexte dans la lecture (position et bases adjacentes).

4.1.4 Détection de variants

À ce stade, les données sont prêtes pour la recherche de variants. Cette recherche se heurte essentiellement à l'impureté des échantillons, aux variations du nombre de copies (CNV, Copy Number Variation), aux erreurs de séquençage ainsi qu'aux régions répétées. La recherche des variants somatiques s'est initialement faite en génotypant les deux échantillons de manière indépendante, puis en soustrayant des variants présents dans l'échantillon tumoral ceux également détectés dans l'échantillon sain. Cette approche peut donner des résultats satisfaisants dans le cas où les échantillons sont purs. Or ce n'est pas la majorité des cas dans l'étude des cancers. Nous passons en revue différents outils permettant de détecter les courtes variations dans le cas général et dans le cas d'échantillons tumoraux. Ces outils sont le plus souvent basés soit sur des heuristiques, soit sur des méthodes probabilistes.

Nous avons répertorié plus d'une vingtaine d'outils pour la détection de variants, parmi lesquels MAQ (Li et al. (2008a)), SAMtools (Li et al. (2009a)), SOAPsnp (Li et al. (2009b)), VarScan [8] (Koboldt et al. (2009)), SNVMix (Goya et al. (2010)), GATK (DePristo et al. (2011), McKenna et al. (2010)), JointSNVMix (Roth et al. (2012)), mutationSeq (Ding et al. (2012)), Cortex (Iqbal et al. (2012)), GeMS [9] (You et al. (2012)) et FreeBayes (Garrison and Marth (2012)), pour ne citer que les plus répandus.

7. Genome Analysis Toolkit
8. URL:http://varscan.sourceforge.net
9. Genotype Model Selection

Une analyse directe des échantillons "pairés" tumoral/sain permet une meilleure distinction entre polymorphismes ou mutations germinales et mutations somatiques et donc permet une nette diminution du nombre de faux positifs (Koboldt et al. (2012), Saunders et al. (2012)). Depuis 2011/2012, des outils dédiés à la détection dans les données de cancer sont développés pour répondre à cette problématique, en permettant une comparaison directe entre un échantillon sain et un échantillon tumoral, comme c'est le cas de VarScan2 (Koboldt et al. (2012)), MutationSeq (Ding et al. (2012)), MuTect (Cibulskis et al. (2013)), Somatic Indel Detector (GATK), Strelka (Saunders et al. (2012)), ou encore SomaticSniper (Larson et al. (2012)). Les outils MuTect et VarScan2 sont les plus utilisés sur les données d'exome pour leur mode de détection "somatic", comme en témoignent les publications telles que Graubert et al. (2011), qui a mis en évidence une mutation dans le gène *U2AF1* chez 9% des patients atteints de syndromes myélodysplasiques. D'autres équipes ont intégré ces comparaisons directes dès 2011 (par exemple Bowne et al. (2011), Matsushita et al. (2012)). MuTect est efficace dans l'analyse d'échantillons tumoraux contaminés par le tissu sain tandis que VarScan2 permet de prendre en compte une contamination tumorale dans l'échantillon sain (Wang et al. (2013b)).

Pour l'analyse des données *WES*, la détection des variants s'est effectuée sur les fichiers Pileup, à partir des fichiers BAM convertis avec SAMtools mpileup version 0.1.18. Lors de la conversion des BAM en mpileup, nous avons supprimé les bases ayant une qualité de base calling < 20, *i.e.* délétion des bases dont la probabilité d'être mal lue est $\geq \frac{1}{100}$ et les lectures ayant une qualité de mapping < 20, *i.e.* délétion des reads dont la probabilité d'être mal alignée est $\geq \frac{1}{100}$. Nous avons utilisé la version 2.3.2 de VarScan2. Dans le mode "somatic", VarScan2 prend en entrée les fichiers d'alignement au format Pileup. Les paramètres suivants sont à fixer : la p-valeur "somatic", la couverture minimum à chaque position pour les deux échantillons, la pureté des deux échantillons, le nombre minimum de reads supportant le variant, le nombre minimum de sens dans lequel doit être lu le variant (1 ou 2), le score minimum de base calling et l'élimination ou non des variants présentant une disproportion supérieure à 90% du nombre de lectures dans l'un des deux sens de lecture. Il faut également définir le pourcentage de variant sous lequel ou au-delà duquel définir une variation comme étant une perte d'hétérozygotie.

Nous considérons toutes les positions ayant été lues au moins une fois. Nous fixons le seuil de qualité des bases à un score $\geq$20, les bases avec un score inférieur à ce seuil ont été éliminées lors de la conversion en fichier Pileup. Nous ne mettons pas de contrainte sur le nombre de lectures de l'allèle variant. Nous n'éliminons pas les variants lus dans une seule direction ou les variants présentant une disproportion dans la répartition des lectures dans les deux directions. Nous avons choisi une p-valeur "somatic" inférieure ou égale à 10^{-4} pour la majorité des échantillons. Cette p-valeur a été augmentée dans le cas de contamination de l'échantillon contrôle et/ou de faible couverture. Nous définissons la pureté des échantillons tumoraux et des échantillons de fibroblastes à 100% et des échantillons de lymphocytes T à 90%. Nous avons défini le seuil d'hétérozygotie à 20% pour la perte de l'allèle variant et à 75% pour la perte de l'allèle normal. Ces seuils dépendent de la contamination des échantillons et de la clonalité de la mutation. L'outil permet uniquement de prendre un seuil "fixe" ; toutefois, le pourcentage dépend de la couverture. Ces seuils permettent de classer les variants en plusieurs catégories mais n'en excluent pas de l'analyse.

Description de la méthode utilisée dans VarScan2

VarScan2 est une méthode basée sur des heuristiques. Il classe les variants comme mutations germinales, mutations somatiques, perte d'hétérozygotie (LOH, Loss Of Heterozygoty) et une catégorie indéterminée. Il est possible de ne conserver que les mutations somatiques par exemple, suivant la problématique. Nous avons gardé toutes les catégories et utilisé la p-valeur somatique

pour supprimer les variants constitutifs car nous avons remarqué des variants mal catégorisés à cause de contamination des échantillons. VarScan2 fournit en sortie deux fichiers contenant les *SNP* et les *INDEL*.

VarScan2, en mode de détection "somatic", compare les génotypes des deux échantillons pour déterminer s'ils sont ou non différents. Le principe est décrit dans l'algorithme 1.

Algorithme 1 : Comparaison des génotypes des 2 échantillons avec VarScan2

```
si l'échantillon tumoral correspond à l'échantillon contrôle alors
    si l'échantillon tumoral et l'échantillon contrôle correspondent à la référence alors
        Définir Référence;
    sinon l'échantillon tumoral et l'échantillon contrôle ne correspondent pas à la référence
        Définir Mutation germinale;
    fin
sinon l'échantillon tumoral ne correspond pas à l'échantillon contrôle
    Calculer la différence de fréquence allélique par FET;
    si la différence est significative (somatic-p-value < seuil) alors
        si l'échantillon contrôle correspond à la référence alors
            Définir Mutation somatique;
        sinon si l'échantillon contrôle est hétérozygote alors
            Définir LOH;
        sinon l'échantillon tumoral et l'échantillon contrôle sont variants, mais différents
            Définir Indéterminé;
        fin
    sinon la différence n'est pas significative
        Définir Mutation germinale;
    fin
fin
```

Deux tests exacts de Fisher (FET) sont réalisés afin d'attribuer un statut à chaque variant. Celui que nous avons utilisé a pour hypothèse nulle H_0 "le génotype de l'échantillon tumoral correspond à celui de l'échantillon normal", *i.e.* la proportion des variants dans l'échantillon tumoral est égale à la proportion des variants dans l'échantillon contrôle. On teste, au risque α (à définir) l'hypothèse nulle du test exact de Fisher réalisé.

Pour les données de *WGS*, la détection des variants s'est effectuée soit sur des fichiers Pileup, avec VarScan v2.3.7, soit sur des fichiers BAM avec SomaticSniper v1.0.3 et Strelka v1.0.14. Nous avons commencé la détection avec SomaticSniper, un outil développé pour la recherche de variants dans les leucémies (où les échantillons contrôles sont souvent contaminés par du tumoral) où les échantillons sont séquencés à une profondeur de 30x. Cet outil correspondait tout à fait à notre étude. SomaticSniper ne permet pas en revanche la détection d'*INDEL*. De ce fait, nous avons utilisé VarScan2 afin d'établir un catalogue complet des courtes anomalies dans le génome des patients *LMMC*. Nous avons constaté le nombre important d'*INDEL* détectés par VarScan2 et avons décidé de comparer ces *INDEL* à ceux détectés par un autre outil. Nous avons opté pour Strelka, un outil basé sur une approche différente des approches classiques.

Description de la méthode utilisée dans SomaticSniper

SomaticSniper (Larson et al. (2012)) est une méthode probabiliste. La détermination du génotype est basée sur le calcul de la vraisemblance de celui-ci et utilise le théorème de Bayes. Le

calcul de la vraisemblance est réalisé à chaque base pour chaque génotype possible. Ce calcul est basé sur les scores de qualité et le nombre de lectures à un site *SNP* donné. Dans l'approche Bayésienne, la vraisemblance est associée à une probabilité *a priori* et leur produit donne la probabilité *a posteriori* recherchée. Le génotype avec la plus grande probabilité *a posteriori* est choisi. Le ratio entre la plus grande et la deuxième plus grande probabilité *a posteriori* peut servir à mesurer l'incertitude du génotype défini.

SomaticSniper compare les vraisemblances des génotypes des échantillons tumoraux et normaux. Pour détecter les mutations somatiques, l'algorithme calcule le score pour qu'un site ne soit pas somatique de la manière suivante : étant donnés un échantillon tumoral T, un échantillon normal N et les génotypes G, le score somatique S est calculé ainsi :

$$S = -10log_{10}\{\sum_{G_i=0}^{9} \frac{p(T|G_i)p(G_i)p(N|G_i)p(G_i)}{\sum_{G_j=0}^{9} p(T|G_j)p(G_j)\sum_{G_k=0}^{9} p(N|G_k)p(G_k)}\} \tag{4.1}$$

La vraisemblance du génotype est $p(D|G_i)$, où D représente l'un des deux échantillons. $p(D|G_i)$ est l'un des 10 génotypes diploïdes possibles : AA,AC,AG,AT,CC,CG,CT,GG,GT,TT. La vraisemblance du génotype est calculée en utilisant l'algorithme MAQ et la probabilité *a priori* $p(G_1)$ est donnée par :

$$p(G_1) = \begin{cases} \theta & \text{Cas 1} \\ \frac{\theta}{2} & \text{Cas 2} \\ \theta^2 & \text{Cas 3} \\ 1 - \sum_{k=0}^{9} p(G_k)p(G_k \neq G_R) & \text{Cas 4} \end{cases}$$

où θ est le taux attendu de mutations hétérozygotes dans la population d'intérêt (fixée ici à 0.001 pour les échantillons humains) et G_R est la base référente à la position d'intérêt. Le cas 1 représente le cas où le génotype est hétérozygote et partage un allèle avec la référence. Le cas 2 représente le cas où le génotype est homozygote variant. Le cas 3 représente le cas où le génotype est hétérozygote, sans allèle commun avec la référence. Le cas 4 représente le cas où le génotype est homozygote référent.

L'équation (4.1) est équivalente à comparer les probabilités des deux mutations comme des mutations germinales indépendantes. La corrélation entre les deux échantillons n'est donc pas prise en compte explicitement. Puisque les deux échantillons proviennent d'un même individu, une modélisation prenant en compte la dépendance de chacun des échantillons avec l'autre est :

$$S = -10log_{10}\{\frac{\sum_{i=0}^{9} p(T|H_i)p(N|G_i)p(H_i|G_i)p(G_i)}{\sum_{j=0}^{9}\sum_{k=0}^{9} p(N|G_k)p(T|H_j)p(H_j|G_k)p(G_k)}\} \tag{4.2}$$

où G est le génotype dans l'échantillon normal et est défini comme précédemment et H est le génotype dans l'échantillon tumoral. $p(H_m|G)$ prend en compte la probabilité *a priori* d'une mutation somatique μ, pour un génotype dans l'échantillon normal G donné de la manière suivante :

$$p(H_m|G) = \begin{cases} \mu & \text{si } H_m \text{ a un allèle commun avec G} \\ \mu^2 & \text{si } H_m \text{ n'a pas d'allèle commun avec G} \\ 1 - \sum_{n=0}^{9} p(H_n|G)p(H_m \neq G) & \text{si } H_m \text{ égal G} \end{cases}$$

Le score somatique se déduit de produits de la vraisemblance des génotypes et de la probabilité *a priori* d'une mutation somatique $p(H_m|G)$. Le détail des calculs est fourni par Larson et al.

(2012).

Nous avons ensuite appliqué VarScan2 en choisissant une p-valeur somatique moins stringente (10^{-2}) qu'avec les données d'exome car la couverture est nettement moins importante (de 3 à 5 fois). Enfin, nous avons utilisé Strelka, qui détecte aussi bien les *SNV* que les *INDEL*.

Description de la méthode utilisée dans Strelka

Strelka (Saunders et al. (2012)) est un workflow de détection de variants somatiques basé sur une approche Bayésienne. Le modèle permet, d'après les auteurs, de maintenir la sensibilité de détection malgré une contamination de l'échantillon tumoral importante, sans pour autant nécessiter une estimation de la pureté. Le workflow de Strelka schématisé figure 4.3, commence par la détection des *INDEL* potentielles. Les lectures autour des *INDEL* candidats détectés sont réalignées. La détection somatique se fait sur les lectures réalignées. L'approche consiste à modéliser les fréquences alléliques plutôt que les génotypes diploïdes. Après sélection des variants présentant un génotype homozygote égal à la référence dans l'échantillon contrôle, les variants sont catégorisés en deux ensembles, tier 1 où les critères sont stringents et tier 2 où seule une qualité de mapping ≥ 5 est requise. Enfin, quelques filtres post détection sont appliqués sur les variants détectés.

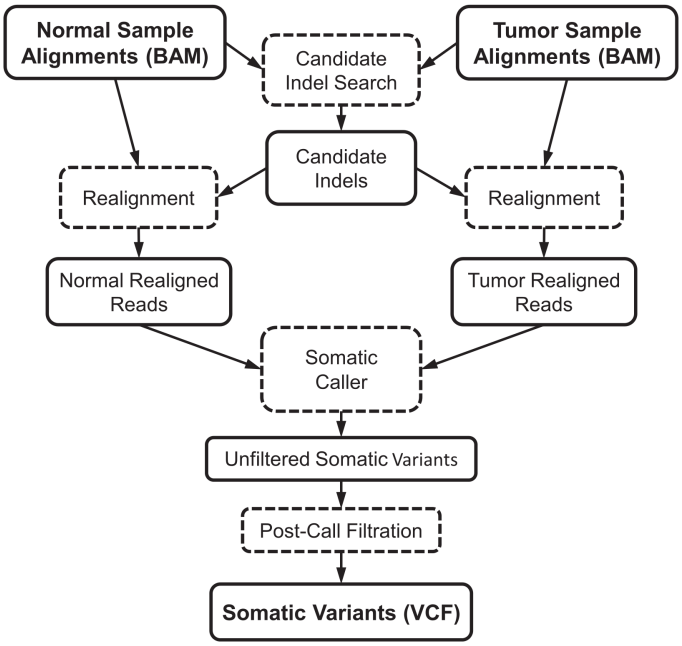


FIGURE 4.3 – *Workflow du variant caller Strelka (Saunders et al. (2012)). Le processus commence avec des fichiers BAM triés desquels les duplicats de PCR ont été marqués pour les deux échantillons. Il s'achève avec l'établissement des SNV et INDEL somatiques*

Que ce soit pour les *SNV* ou les *INDEL*, le modèle approxime une probabilité *a posteriori* des fréquences alléliques tumorales et normales :

$$p(f_t, f_n|D) \propto p(D|f_t, f_n)p(f_t, f_n)$$

où f_t, f_n et D représentent les fréquences alléliques tumorales, normales et les données de séquençage des deux échantillons.

La vraisemblance se calcule comme le produit des vraisemblances indépendantes des fréquences alléliques spécifiques à chacun des échantillons :

$$p(D|f_t, f_n) = p(D_t|f_t)p(D_n|f_n)$$

où D_t et D_n représentent les données de séquençage des échantillons tumoraux et normaux et

$$p(D_x|f_x) = \prod_{b \in D_x} \sum_{a \in A} p(b|a)p(a|f_x)$$

où A est l'ensemble des 4 allèles possibles, p(b | a) est la probabilité d'observer la base b sachant que la vraie base est a, p(a | f_x) est la probabilité d'échantillonnage de la base a, une valeur égale à la fréquence de l'allèle a dans f_x. Étant donné la probabilité d'erreur de séquençage e,

$$p(b|a) = \begin{cases} 1-e & \text{si } b = a \\ \frac{e}{3} & sinon \end{cases}$$

La probabilité *a priori* des fréquences alléliques tumorales et normales p(f_t, f_n) traduit l'idée que l'échantillon normal est un mixte entre des variations germinales diploïdes et un bruit, alors que l'échantillon tumoral est un mixte entre l'échantillon normal et des variations somatiques.

La probabilité *a priori* de la fréquence allélique de l'échantillon normal s'écrit donc : $p(f_n) = p_{diploid}(f_n)(1-\mu) + p_{noise}(f_n)\mu$.

Le produit de la vraisemblance des fréquences alléliques et de la probabilité *a priori* donne la probabilité *a posteriori* recherchée. Cette probabilité *a posteriori* est utilisée pour calculer la probabilité d'obtenir un variant somatique étant donné le statut des échantillons S=(f_t,f_n) :$f_t \neq f_n$

$$p(S|D) = \int_{f_t, f_n} I_s(f_t, f_n)p(f_t, f_n|D)$$

où $I_s(f_t, f_n)$ est la fonction indicatrice de l'état somatique. Cette probabilité de détection d'un variant somatique ne permet pas de distinguer les types de variants somatiques. C'est pourquoi la dernière étape consiste à associer à p(S | D) le génotype de l'échantillon sain :

$$p(S, G_n|D) = p(S|D)p(G_n|D_n)$$

Le génotype de l'échantillon normal $p(G_n|D_n)$ qui nous intéresse pour la recherche de variants somatiques est restreint au cas d'homozygotie égale à la référence. Tous les calculs sont décrits dans les données supplémentaires de Saunders et al. (2012).

Les paramètres choisis lors de l'utilisation de SomaticSniper et Strelka ont été définis dans les méthodes de la publication.

Nous avons utilisé trois méthodes pour détecter les *SNV* et deux méthodes pour les *INDEL*. Après avoir appliqué les mêmes filtres sur les variants détectés par chaque méthode (les filtres sont décrits dans les méthodes de la publication), nous avons comparé les trois ensembles de *SNV* et deux ensembles d'INDELs (figure 4.4).

Afin de déterminer s'il convenait de poursuivre avec les variants détectés par tous les outils ou au contraire par deux ou trois outils, nous avons visualisé 100 variants choisis aléatoirement dans les quatre ensembles de variants formés par l'intersection de deux des outils. Ce critère est loin d'être parfait mais a les mérites d'être relativement rapide et de ne pas occasionner de coûts. Ainsi, 17/100 des variants détectés par SomaticSniper et VarScan2 semblaient vrais, contre 11/100 par SomaticSniper et Strelka et 24/100 par Strelka et VarScan2. Pour les *INDEL*, 63/100 des variants détectés par Strelka et VarScan2 semblaient vrais. Néanmoins, 89/100 des variants détectés par

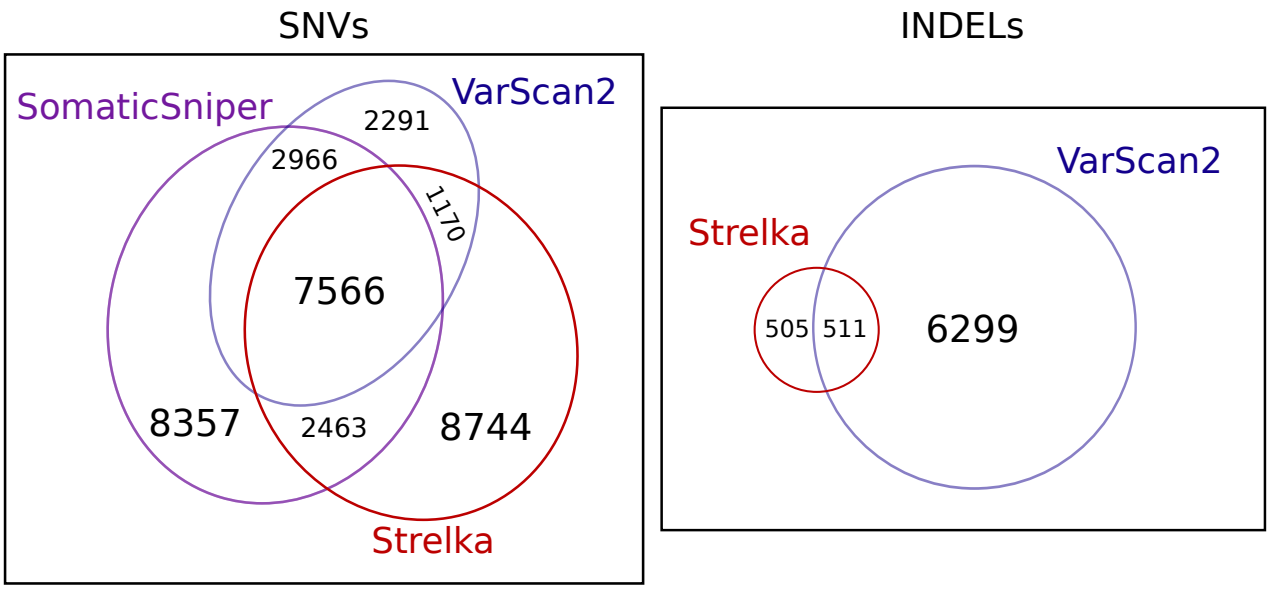


FIGURE 4.4 – *Comparaison des variants détectés dans le génome par plusieurs algorithmes de détection*

SomaticSniper, Strelka et VarScan2 semblaient vrais. Aussi avons-nous décidé de poursuivre les analyses avec les variants détectés par les trois méthodes pour les *SNV* et deux méthodes pour les *INDEL*, en gardant en tête qu'il reste des faux positifs et que nous avons éliminé de vraies variations.

4.1.5 Annotation de variants

Une fois la détection réalisée, les variants peuvent être annotés. Les outils disponibles sont une fois encore nombreux. Les outils permettent d'aller chercher les informations disponibles dans les bases de données publiques de manière automatique pour chaque variant.

Bases de données publiques

Les variants peuvent être annotés selon leur fréquence, la conservation de leur position ou leur effet essentiellement. Nous expliquons dans cette partie quels sont les critères les plus couramment étudiés.

Tout d'abord, les projets NHLBI GO Exome Sequencing Project (ESP) (Server (2012)), 1000 Génomes (Kaiser (2008)) et HapMap (Gibbs et al. (2003)) fournissent la fréquence du variant dans la population générale. Le but de *ESP* est de découvrir de nouveaux gènes et mécanismes contribuant à des pathologies cardiaques, pulmonaires et hémopathiques en utilisant le séquençage *NGS* d'exomes dans plusieurs populations humaines. Les données et résultats sont partagés avec la communauté scientifique pour améliorer le diagnostic et le traitement de ces pathologies. Le projet 1000 Génomes a pour but de capter la variation génétique chez l'humain. Plus précisément, il vise à trouver la plupart des variants génétiques ayant une fréquence d'au moins 1% dans la population générale. Le projet est destiné à couvrir 2500 échantillons à une couverture de 4x. Le projet HapMap établit un catalogue des variations génétiques les plus fréquentes chez l'humain. Il est issu d'un partenariat de scientifiques et d'agences de financement qui vise à aider les chercheurs à trouver des gènes associés à des pathologies humaines dans le but d'aboutir à de nouvelles méthodes de prévention, de diagnostic et de traitement des maladies.

La base de données des polymorphismes (dbSNP, Single Nucleotide Polymorphism database) (Sherry et al. (2001)) a été conçue pour recevoir les soumissions et soutenir la recherche d'un

important panel de problèmes biologiques : alignement de séquences, médicaments, études d'association et études d'évolution. Les polymorphismes répertoriés incluent les *SNP*, les *INDEL*, les rétrotransposons et les séquences microsatellites. La version 129 est considérée comme la dernière version "propre", ne contenant que des polymorphismes et pas de mutation.

Un autre élément important est le degré de conservation du nucléotide altéré à travers différentes espèces. PhyloP [10] (Pollard et al. (2010)), PhastCons [11] (Siepel et al. (2005)) et GERP++ (Davydov et al. (2010)) par exemple ont pour but d'identifier des éléments conservés au fil de l'évolution.

Enfin, les bases COSMIC [12] (Bamford et al. (2004)) et ClinVar (Landrum et al. (2014)) répertorient les mutations somatiques observées dans tous les types de cancers et dans les syndromes bénins pour ClinVar.

Algorithmes de prédiction

Des algorithmes ont été développés afin de prédire la conséquence fonctionnelle d'un changement d'acide aminé sur une protéine lorsque celle-ci est associée à une perte de fonction. Les deux plus répandus sont SIFT [13] (Ng and Henikoff (2003), Sim et al. (2012)) et PolyPhen-2 [14] (Ramensky et al. (2002)). Il y a également LRT [15] (Chun and Fay (2009)). Chun and Fay (2009) ont observé dans l'étude de trois génomes humains que 76% des prédictions délétères sont fournies par un seul des trois algorithmes et que seulement 5% des prédictions délétères sont communes aux trois algorithmes. MutationTaster (Schwarz et al. (2010)) intègre les données de différentes bases de données biomédicales : les analyses comprennent les changements de conservation, les changements de sites d'épissage, les pertes de caractéristiques de protéines et les changements qui pourraient affecter la quantité d'*ARNm*. Ces outils doivent être maniés avec précaution et uniquement à titre indicatif puisque les prédictions ne sont pas toujours correctes.

Outils d'annotation

Les outils d'annotation permettent d'annoter les *SNP* et *INDEL* sur la position génomique, le type de mutations et peuvent s'associer aux bases précédemment citées. Certains outils permettent d'annoter les *SNP* comme F-SNP (Lee and Shatkay (2008)) et VARIANT [16] (Medina et al. (2012)). F-SNP prédit l'effet fonctionnel de *SNP* à partir de 16 outils bioinformatiques et bases de données. D'autres annotent aussi bien les *SNP* que les *INDEL* comme SNPnexus (Chelala et al. (2009), VAAST [17] (Yandell et al. (2011)) ou VariantAnnotation (Obenchain et al. (2014)). Anntools (Makarov et al. (2012)), Annovar (Wang et al. (2010)) et SnpEff (Cingolani et al. (2012)) annotent les *SNV*, *INDEL* et *CNV*. Ces deux derniers outils sont les plus utilisés.

Nous avons réalisé les annotations avec Annovar. Nous avons déterminé la position du variant : exon, intron, région intergénique, régions non traduites 5' et 3', *ARN* non codant, zone "d'épissage". Les variants se trouvant dans les exons sont définis comme des *INDEL* décalant ou non le cadre de lecture ((non)frameshift), des *SNV* synonymes ou non synonymes et des pertes ou gains de codons stop. Les variants se trouvant dans les zones "d'épissage" se trouvent jusqu'à 10 bases des extrémités des exons. Nous avons utilisé plusieurs bases de données pour

10. phylogenetic P-values
11. phylogenetic analysis with space/time models, Conservation
12. Catalogue Of Somatic Mutations In Cancer
13. Sort Intolerant From Tolerant
14. Polymorphism Phenotyping v2
15. Likelihood Ratio Test
16. VARIant ANalysis Tool
17. the Variant Annotation, Analysis and Search Tool

décrire les variants : dbSNP (versions 129 et 138), SIFT, PolyPhen, ESP6500, 1000G et PhyloP pour caractériser chaque variant en détails. Nous avons utilisé de plus COSMIC et ClinVar pour les données de séquençage ciblé et les données de génome.

Pour l'étude de l'exome, les variants se trouvant dans les exons ont été conservés, à l'exception des variants synonymes. Les kits de capture d'exome couvrant les régions adjacentes aux exons, il est possible de détecter des variants non exoniques. Notre liste finale contient insertions et délétions, qu'elles décalent ou pas le cadre de lecture, des pertes ou gains de codons stop et les *SNV* non synonymes. Nous avons supprimé les variants se trouvant dans dbSNP129, quelle que soit leur prédiction fonctionnelle. Les variants retenus sont ceux ayant passé tous ces filtres. Pour l'étude du génome, tous les variants sauf ceux répertoriés dans dbSNP129 ont été conservés.

4.1.6 Analyse du nombre de copies et de pertes d'hétérozygotie

La possibilité de détecter les *CNV* et les *LOH* à partir de données *NGS* de génome ou d'exome étend les applications de ce séquençage utilisé essentiellement pour détecter mutations et *INDEL*. Plusieurs outils ont ainsi été développés ces dernières années pour accéder à ces informations en utilisant directement les données générées par *NGS*. Jusqu'à récemment, l'étude des *CNV* se faisait uniquement par hybridation génomique comparative (CGH (Comparative Genomic Hybridization) microarray ou puce) (Pinkel et al. (1998)) et l'étude des *LOH* par SNP-arrays (Single Nucleotide Polymorphism genotyping arrays). L'hybridation génomique comparative apparue en 1992 permet la détection d'événements d'au moins 5Mb.

Les premiers outils développés pour les données *NGS* visaient à détecter les anomalies dans des données *WGS*. Nous pouvons citer SeqSeq (Chiang et al. (2008)), EWT [18] (Yoon et al. (2009)), HMMcopy (Lai and Ha (2012)) (package R) et cn.MOPS [19] (Klambauer et al. (2012)). Certaines de ces méthodes font des hypothèses qui les rendent inutilisables sur des données d'exome. Par exemple, l'hypothèse d'une distribution aléatoire et non biaisée des lectures, de telle façon que la profondeur de lecture peut être modélisée par une distribution uniforme le long du génome, une déviation du bruit de fond indiquant la présence de CNV. Cette hypothèse de distribution des lectures n'est plus vérifiée dans le contexte d'une capture d'exome. Aussi, des outils permettant de traiter spécifiquement ou en complément des données d'exome ont été développés. Pour l'étude du nombre de copies, VarScan et exomeCopy (Love et al. (2011)) (package R) peuvent être utilisés avec ou sans contrôle. ExomeCNV (Sathirapongsasuti et al. (2011)) permet de détecter à la fois les *CNV* et les *LOH*, en utilisant ou non un échantillon contrôle. C'est également le cas de Control-FREEC [20] (Boeva et al. (2012)) qui a été adapté pour l'étude des données *WES*.

Nous avons utilisé Control-FREEC, version v6.4, avec les options suivantes : window=500, breakPointThreshold=0.5, noisyData=TRUE et minimalCoveragePerPosition=5, les autres paramètres étant par défaut. Nous avons conservé les événements d'au moins 5Mbp fournis dans les fichiers _CNVs pour lesquels l'incertitude du génotype prédit était inférieure ou égale à 10.

L'analyse implémentée dans Control-FREEC se fait en trois parties : le calcul et la segmentation des profils du nombre de copies, le calcul et la segmentation des fréquences alléliques variantes et la prédiction du génotype de chaque segment détecté.

La détection du nombre de copies se fait en quatre étapes :

1. Calcul des profils de nombre de copies

18. Event-wise testing
19. Mixture Of PoissonS for discovering Copy Number variations in next generation sequencing data
20. Control-FREE Copy number and allelic content caller

Pour cela, on compte le nombre de lectures dans des fenêtres non chevauchantes.

2. Normalisation des profils bruts de nombre de copies
   Pour la normalisation, on effectue une régression par la méthode des moindres carrés entre le nombre de lectures observé dans l'échantillon tumoral et celui observé dans l'échantillon contrôle. Le degré du polynôme est 1, 2 ou 3 le plus souvent. Plusieurs hypothèses sont faites :
   — la ploïdie est donnée
   — le nombre de lectures observé dans les régions à P-copies peut être modélisé par un polynôme du nombre de lectures observé dans l'échantillon de contrôle
   — le nombre de lectures observé dans une région avec un nombre de copies altéré est proportionnel au nombre de lectures dans les régions à P-copies
   — l'intervalle du nombre de lectures observé dans l'échantillon de contrôle dans les régions de principale ploïdie doit inclure l'intervalle de toutes les lectures dans l'échantillon de contrôle.
3. Segmentation des profils normalisés
   L'objectif est de déterminer le nombre de segments et la position des points de cassure qui expliquent au mieux et le plus simplement possible les données. Chaque chromosome est segmenté indépendamment.
   La segmentation se fait par une méthode basée sur l'algorithme LASSO (Least Absolute Shrinkage and Selection Operator).
4. Analyse des profils segmentés
   Les segments sont définis comme régions de pertes ou gains et le nombre de copies dans ces régions est établi. Le nombre de copies est l'arrondi à l'entier le plus proche du produit du nombre de lectures normalisé et de la ploïdie. Les régions centromériques et télomériques étant sujettes à de nombreux artefacts, elles se voient attribuées le nombre de copies des régions voisines.

Pour l'analyse des *LOH*, on utilise les *SNP* répertoriés. Seuls les *SNP* hétérozygotes fournissent de l'information dans le cas d'anomalies et sont donc utilisés. On calcule la fréquence allélique de chacun d'eux. On calcule la médiane des écarts entre la fréquence allélique observée et la fréquence allélique attendue (50%) dans chaque fenêtre analysée. Les profils de ces médianes sont segmentés avec le même algorithme de segmentation que celui utilisé pour les profils des nombres de copies.

4.1.7 Les spécificités du séquençage ciblé

Le séquençage à très haut débit ciblé sur de courtes régions permet de séquencer à une très grande profondeur ($\gg$1000x) pour déceler des mutations présentes dans de petits sous-clones. De cette manière, il est possible de capter l'hétérogénéité tumorale et d'essayer d'établir l'ordre d'apparition des mutations. Il est également utilisé pour confirmer des variants détectés par *WES* ou *WGS*. Les variants détectés dans nos données d'exome ont été confirmés par séquençage ciblé avec l'Ion Torrent PGM à une profondeur de 759x pour définir le taux de faux positifs généré par ce processus. Ceci fournit un indicateur de la qualité de la méthode. Nous avons confirmé 71% des variants, 9% n'ont pas été lus et 20% n'ont pas été confirmés. Le séquençage ciblé permet aussi d'analyser de courtes régions (hotspot, gène), comme nous l'avons fait pour les gènes candidats étudiés en MiSEQ à une profondeur de 710x.

4.1.8 Les spécificités du séquençage d'exome

Le *WES* permet l'étude des régions codantes du génome à une couverture de 100-200x en moyenne. Le défaut du *WES* est que la couverture n'est pas uniforme et une partie des bases

codantes est peu couverte. Un faible pourcentage des régions codantes n'est pas lu, mais ceci tend à s'améliorer au fur et à mesure des versions des kits de capture. Ce séquençage semble un bon compromis entre l'étude de gènes candidats (biaisée par les connaissances *a priori*) et l'étude du génome entier, qui reste cher et génère de volumineuses données, dont la majorité des mutations sont de significations inconnues.

4.1.9 Les spécificités du séquençage de génome

Le *WGS* se fait généralement à une faible profondeur ($\simeq$30x) en raison des coûts encore assez élevés et de la quantité des données générées. Quelques récents projets atteignent des couvertures de 60-80x. Contrairement aux données d'exome ou de séquençage ciblé, il n'est pas envisageable de rechercher des mutations dans des sous-clones. En revanche, la couverture est uniforme le long du génome, beaucoup plus que dans l'exome et permet une recherche plus aisée des *CNV* et des *LOH*. Les télomères et centromères sont encore mal gérés, soit non séquencés, soit séquencés à une très grande profondeur (plusieurs milliers de reads).

Le génome se compose de nombreuses régions répétées (environ 50% du génome), contenant les microsatellites. De plus, il existe des régions très polymorphiques d'un individu à l'autre, comme les régions HLA ou les récepteurs olfactifs. L'alignement des lectures relativement courtes (100 bases) venant de ces régions sur le génome de référence est difficile et génère de nombreux alignements erronés. De nombreuses variations apparaissent ainsi de manière artefactuelle. La gestion des régions répétées est pour cette raison un vrai défi. Pour se soustraire à ces nombreux artefacts, il est possible de ne pas étudier ces régions. Suivant les sources, les régions répétées à éliminer représentent de quelques pourcents à 45% du génome. Dans les résultats présentés dans la suite, nous avons éliminé 45% du génome.

Les données de génome ont l'avantage de permettre la recherche de variations structuales, *i.e.* des variations affectant quelques dizaines à quelques milliers de bases : translocations, inversions, grandes insertions et délétions. Pour détecter de telles variations, des outils spécifiques ont été développés : BreakDancer (Chen et al. (2009)), Pindel (Ye et al. (2009)), DELLY (Rausch et al. (2012)), SVDetect (Zeitouni et al. (2010)),... Nous avons entrepris cette étude en collaboration avec le Centre National de Génotypage où les échantillons de génomes ont été séquencés. L'analyse est encore en cours.

Les données de génome ont également l'avantage de permettre une recherche non biaisée (contrairement aux données *WES*) de signatures mutationnelles. Une signature représente l'empreinte laissée par un processus connu ou inconnu à l'origine de mutations chez un individu ou plus largement chez un ensemble d'individus ayant le même cancer. On étudie les six différents changements de bases possibles en reportant leurs bases avoisinantes. Alexandrov et al. (2013b) ont proposé une méthode d'analyse pour détecter les signatures de chaque individu. Un même individu peut être porteur de plusieurs signatures et chaque signature peut expliquer un nombre variable des mutations de l'individu. En collaborant avec Alexandrov, nous avons pu identifier 3 signatures chez nos 17 patients. Deux des trois signatures (signatures 1 et 5 dans la publication d'Alexandrov et al. (2013a)) sont présentes chez chaque patient, tandis que la troisième signature n'est trouvée que chez 2 des patients et n'a jamais été observée par le groupe d'Alexandrov. Nous ne savons pas à quel processus correspond cette signature. Les deux patients en question présentent une double pathologie, mais la deuxième pathologie n'est pas la même dans les deux cas : *LMMC*/histiocytose langerhansienne et *LMMC*/lymphome. Ces cas sont décrits dans la figure 2 de la publication.

4.1.10 Séquençage et analyse du niveau de méthylation de l'ADN

Le séquençage et l'analyse des profils de méthylation des patients non traités, traités et répondeurs et traités et stables ont été réalisés par l'équipe de Maria Figueroa (University of Michigan Medical School, Department of Pathology, Ann Arbor, Michigan). Nous nous contentons de donner les principes de ce séquençage et les grandes lignes de l'analyse.

Il existe essentiellement deux principes pour capturer le niveau de méthylation d'un génome : des approches de fragmentation aléatoire de l'*ADN* et des approches de digestion par l'enzyme de restriction MspI (Lister and Ecker (2009)). Nous avons utilisé l'Extended Reduced Representation Bisulfite Sequencing (ERRBS) (Garrett-Bakelman et al. (2015)) qui est basée sur une digestion MspI. Cette méthode est une amélioration du Reduced Representation Bisulfite Sequencing (RRBS) (Meissner et al. (2005)). Le *RRBS* permet de couvrir la majorité des promoteurs des gènes et des îlots CpG, mais la couverture des amas d'îlots CpG et autres régions introniques est limitée (Garrett-Bakelman et al. (2015)). L'*ERRBS* permet de capter plus d'îlots CpG et d'augmenter la couverture de toutes les régions génomiques ciblées (Garrett-Bakelman et al. (2015)).

Les étapes précédant le séquençage sont semblables à celles d'un séquençage classique de l'*ADN*, à ceci près qu'une étape de conversion bisulfite est nécessaire avant les *PCR*. Les cytosines sont converties en thymine. Les cytosines méthylées ne peuvent être converties. L'analyse débute de manière classique, par un contrôle qualité et si besoin, par la suppression partielle ou complète de lectures, suivie d'un alignement des lectures sur le génome de référence converti. La taille des lectures a été 50 bp. Bismark (Krueger and Andrews (2011)) a été utilisé pour l'alignement. Le taux d'alignement est classiquement de l'ordre de 50-70%. Dans notre cas, il a été de 62% en moyenne. La couverture moyenne a été 63x. A chaque position est défini un score de méthylation, représentant le pourcentage de conversion bisulfite des cytosines (les cytosines non méthylées) et le pourcentage de non-conversion bisulfite des cytosines (les cytosines méthylées). L'analyse des régions différentiellement méthylées a été faite avec le package R methylKit (Akalin et al. (2012)).

4.2 Analyse de séquences d'ARN

Le *RNASeq* capture les séquences des *ARN* transcrits dans un tissu à un moment précis. Les questions posées sont essentiellement la recherche de dérégulations de l'expression génique entre différentes conditions, d'épissages alternatifs ou de transcrits de fusion. Depuis 2 ans, la recherche de variations dans les données de *RNASeq* se développe dans le but de s'affranchir de la génération de données *WES*. Il serait ainsi possible de rechercher les mutations affectant les régions exprimées sans générer de données supplémentaires.

4.2.1 Alignement de séquences d'ARN sur une référence

L'alignement de séquences d'*ARN* diffère de celui des données d'*ADN* puisque les transcrits sont constitués majoritairement d'exons et pas forcément de tous les exons d'un gène. L'alignement doit donc permettre de grandes délétions par rapport au génome de référence. Plusieurs outils d'alignement de séquences d'*ARN* ont été développés, TopHat (Trapnell et al. (2009)), CRAC (Philippe et al. (2013)) et surtout TopHat2 (Kim et al. (2013)) sont les plus largement utilisés. Dans cette étude, nous avons aligné les lectures sur un transcriptome de référence avec TopHat2. Les lectures non alignées ont été sujettes à un alignement sur le génome de référence. Enfin, les lectures toujours non alignées ont été fragmentées en des régions de 25 bases et ces petits fragments ont été alignés sur le génome de référence.

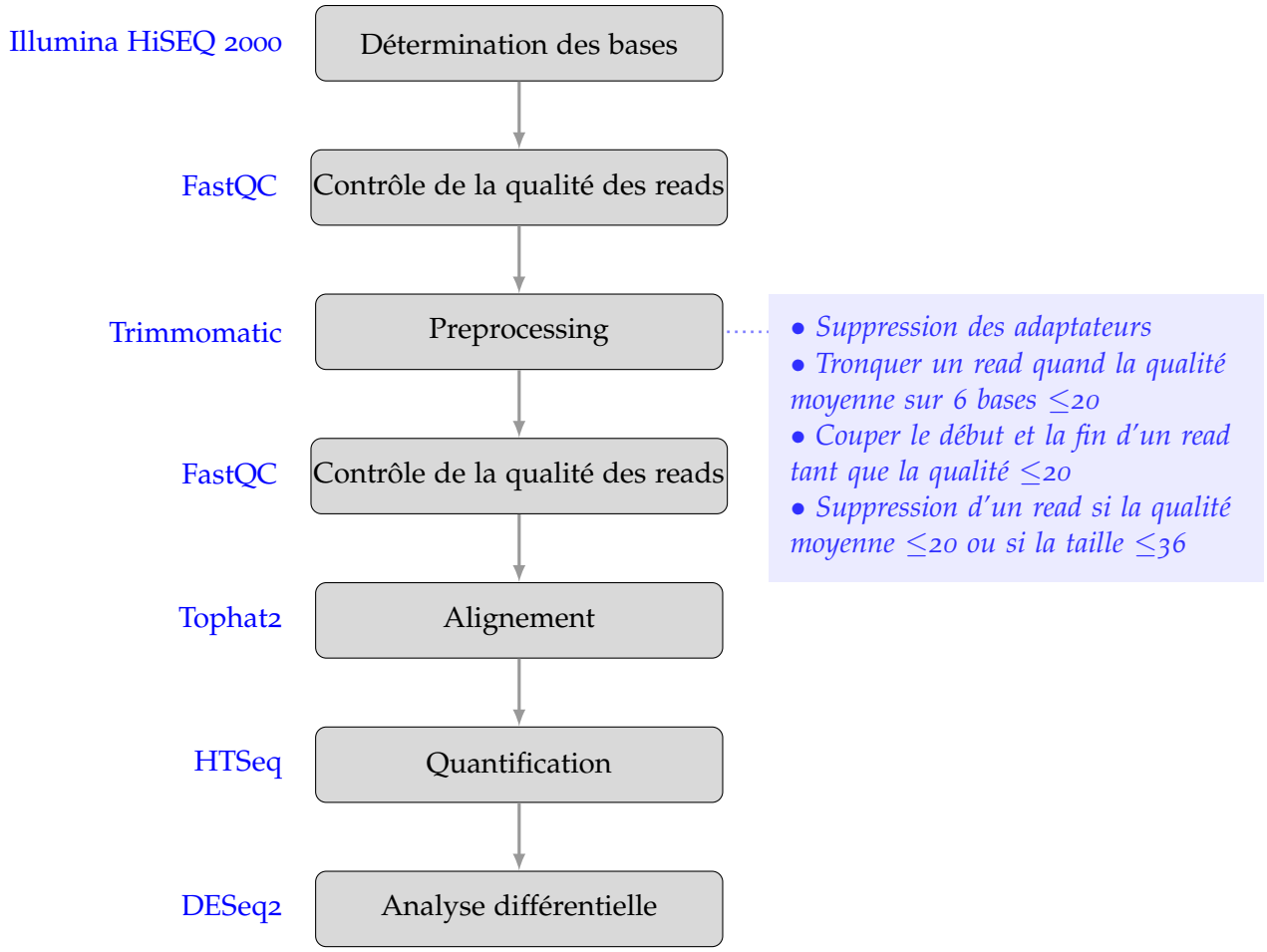


FIGURE 4.5 – *Analyse réalisée sur les données RNASeq pour l'étude des dérégulations géniques*

4.2.2 Analyse d'expression différentielle

Les étapes réalisées pour l'analyse d'expression différentielle sont schématisées figure 4.5.

Comptage des reads

Pour rechercher des anomalies d'expression entre plusieurs conditions, la première étape est la quantification du niveau d'expression de l'ensemble des gènes de chaque échantillon. Il faut pour cela procéder au comptage du nombre de lectures dans chaque entité (gène, transcrit, isoforme ou encore exon) pour chaque échantillon. Dans cette étude, nous avons recherché les gènes différentiellement exprimés. HTSeq-count (Anders et al. (2014)) permet de comptabiliser le nombre de lectures s'alignant sur une certaine entité, les gènes dans notre cas.

Normalisation des comptages

La deuxième étape consiste en la normalisation des données de comptage obtenues. Cette étape est essentielle car elle permet de rendre les échantillons comparables même si la quantité de matériel déposé n'est pas la même pour tous les échantillons, sans quoi la majorité des gènes serait moins (ou plus) fortement exprimée dans un échantillon donné. Plusieurs méthodes ont été développées pour répondre à cette problématique : RPKM [21] (Trapnell et al. (2012)), GenomeGraphs (Durinck et al. (2009)) (la méthode développée Upper Quartile (Bullard et al. (2010)) est plus connue que le package R), DESeq (Anders and Huber (2010)), edgeR (Robinson and Oshlack (2010)), ... Ces méthodes ont été comparées par Dillies et al. (2012). Les méthodes recommandées

21. Reads Per Kilobase per Million mapped reads

sont DESeq et edgeR. Dans ces deux méthodes entre autres, le paramètre de normalisation ne tient pas compte de la longueur du gène puisque celle-ci ne varie pas lorsque l'on compare l'expression d'un gène dans deux conditions. L'outil DESeq a été amélioré dans DESeq2 (Love et al. (2014)).

Nous présentons uniquement le modèle utilisé dans DESeq2, qui a été utilisé pour les analyses. Pour l'estimation du facteur de normalisation, la méthode de la médiane des ratios est utilisée. On suppose que la majorité des gènes n'est pas différentiellement exprimée. Pour un gène d'un échantillon donné, on compare sa valeur d'expression par rapport aux valeurs des autres échantillons. On répète cette opération pour l'ensemble des gènes. Plusieurs milliers de ratios sont calculés et d'après l'hypothèse selon laquelle la majorité des gènes n'est pas différentiellement exprimée, la médiane de ce ratio doit être égale à 1. Le facteur de normalisation est le nombre réel à appliquer à la médiane pour la rendre égale à 1.

Généralités sur DESeq2

DESeq2 propose un modèle pour l'analyse de l'expression différentielle des gènes. Le modèle statistique compare la quantité de transcrits de chaque gène dans au moins deux conditions en prenant en compte les réplicats biologiques. Le nombre de lectures de chaque gène est proportionnel à l'abondance de ses transcrits. Nous souhaitons comparer ces nombres de lectures entre les conditions biologiques. On teste si, pour un gène donné, la différence observée du nombre de reads des conditions est significativement différente ou non. Le modèle le plus couramment employé suppose que les lectures sont indépendamment et uniformément échantillonnées à partir d'une population d'*ARN*. Ainsi, la proportion de lectures associées à un gène donné reflète la fraction d'*ARN* associée au gène dans la population d'origine. Le nombre de lectures associées à chaque gène suit une loi multinomiale, qui peut être approximée par un ensemble de lois de Poisson. Toutefois, le cadre de la distribution de Poisson était trop restrictif car prédisait des variations plus petites que celles observées dans les données. Le test statistique en résultant ne contrôlait pas le taux de fausse découverte. Pour pallier ce problème, une alternative robuste à la loi de Poisson est utilisée pour modéliser la surdispersion : la loi binomiale négative.

Modèle utilisé dans DESeq2

Les données de comptage sont contenues dans une matrice K avec en ligne, chaque gène i et en colonne, chaque échantillon j. Les valeurs de K_{ij} représentent le nombre de reads alignés sur un gène pour un échantillon donné. Les K_{ij} sont modélisés par un modèle linéaire généralisé (GLM, Generalized Linear Model). Par définition d'un *GLM*, les K_i sont indépendants, la loi des K_i appartient à une famille exponentielle de lois, ici une loi binomiale négative d'espérance μ_{ij} et dispersion α_i et la fonction de lien, ici logarithmique, décrit les variations de $E(K_i)$ en fonction d'un régresseur linéaire : $\log(E(K_i))=x_i\beta$.

$$K_{ij} \mapsto NB(\mu_{ij}, \alpha_i)$$

μ_{ij} représente la quantité d'*ADNc* d'un gène donné et est supposée proportionnelle à la quantité d'*ADNc* observée pour ce gène, pourvu que l'échantillonnage soit correct.

$$\mu_{ij} = s_{ij} q_{ij}$$

$$log_2(q_{ij}) = \sum_r x_{jr} \beta_{ir}$$

où x_{jr} sont les éléments de la matrice de design (indiquant la condition à laquelle appartient chaque échantillon) et β_{ir} la matrice des coefficients.

s_{ij} est le facteur de normalisation. Dans la majorité des cas, un même facteur de normalisation peut être appliqué à tous les gènes d'un même échantillon. Pour rappel, dans le modèle choisi, c'est la méthode des médianes des ratios qui est employée.

$$s_j = median_{i:K_i^R \neq 0} \frac{K_{ij}}{K_i^R}$$

avec

$$K_i^R = (\prod_{j=1}^{m} K_{ij})^{1/m}$$

Le *GLM* retourne les coefficients indiquant le niveau d'expression moyen et le $\log_2$(fold change) entre conditions pour chaque gène. Les *GLM* permettent d'étudier des expériences aux designs plus complexes (comparaisons de plusieurs conditions, d'une sous partie des conditions,...).

Estimation de la dispersion par DESeq2

Le changement majeur entre DESeq et DESeq2 est l'estimation de la dispersion des valeurs d'expression. La nouvelle approche est décrite dans Love et al. (2014). Sans rentrer dans les détails, nous mentionnons juste que le paramètre de dispersion α_i est une fonction de la moyenne. La difficulté pour l'estimer tient dans le faible nombre de réplicats généralement étudiés.

Test statistique utilisé dans DESeq2

Après avoir appliqué le *GLM* à chaque gène, on teste si les coefficients du modèle diffèrent significativement de 0. DESeq2 calcule l'erreur standard de chaque estimation du $\log_2$(fold change). DESeq2 propose le test de Wald ou le test du rapport de vraisemblance. Ils permettent de tester des coefficients seuls ou des combinaisons de coefficients. Les p-valeurs issues du test sont ajustées pour les tests multiples par la procédure de Benjamini-Hochberg dans notre cas.

4.2.3 Variants d'épissage

La recherche d'anomalies d'épissage requiert un nombre de lectures beaucoup plus important (100 à 200 millions de lectures par échantillon d'après les standards de l'Encyclopedia of DNA Elements (ENCODE)) que pour l'étude d'expression différentielle (20 à 30 millions de lectures par échantillon d'après les standards *ENCODE*). Il est essentiel d'avoir de nombreuses lectures couvrant chaque jonction exon-intron pour déceler l'absence ou la présence d'un exon, intron ou transcrit, mais surtout pour déterminer une différence de proportion. Plusieurs outils comme MISO (Katz et al. (2010)) ou cuffdiff2 (Trapnell et al. (2013)) ont été développés afin d'estimer le niveau d'expression de chaque transcrit dans le but de faire une analyse différentielle des isoformes, mais leurs estimations ne sont pas toujours très fiables.

Dans notre recherche d'anomalies d'épissage, nous ne nous sommes pas confrontés au challenge de l'estimation des proportions d'isoformes. Nous nous sommes limités à l'étude des jonctions en utilisant une méthode développée par nos collaborateurs. Ces analyses ont été réalisées par Émilie Chautard, dans l'équipe de Didier Auboeuf spécialisée dans l'étude des événements d'épissage (Centre Léon Bérard, INSERM U1052, CNRS UMR5286, Lyon). Nous nous sommes concentrés sur la détection des sauts d'exon, sauts de plusieurs exons, exons mutuellement exclusifs et sites 3' et 5' alternatifs, sans chercher à établir le niveau d'expression de chaque transcrit. Pour chaque événement est mesuré le Percent Spliced In ($\Delta\psi$), qui représente le pourcentage d'inclusion d'un événement : nombre de lectures supportant l'événement par rapport au nombre de lectures totales. Les événements d'épissage présentant une proportion significativement différente et une différence de $\Delta\psi \geq 20\%$ entre les conditions étudiées sont présentés dans la partie 6.2.

4.2.4 Détection de fusions

De nombreux outils ont été développés pour rechercher les fusions dans les séquences d'*ARN*. Les plus connus sont FusionSeq (Sboner et al. (2010)), ChimeraScan (Iyer et al. (2011)), deFuse (McPherson et al. (2011)), FusionHunter (Li et al. (2011)), FusionMap (Ge et al. (2011)), TopHat2 avec l'option fusion-search et CRAC (Philippe et al. (2013)). A l'heure actuelle, ces analyses détectent de nombreux faux positifs en raison d'alignements incorrects dus aux séquences répétées, *CNV*, pseudogènes, ... Les événements détectés doivent être visualisés puis ceux semblant fiables doivent être confirmés expérimentalement par *PCR*. Nos recherches de fusions avec TopHat2 puis avec CRAC (menées par l'équipe de Thérèse Commes (Institut de Recherche en Biothérapie, INSERM U1040, Montpellier), ne nous ont pas permis d'identifier des candidats intéressants. Nous ne rentrons donc pas dans les détails des méthodes.

Pour la plupart des outils, la recherche de fusions se fait de la manière suivante : les lectures sont alignées sur un transcriptome et/ou génome de référence. Les lectures non alignées ou les paires alignées avec une taille d'insert supérieure à celle attendue sont analysées dans la suite. Les lectures non alignées sont coupées en au moins deux morceaux et chaque partie est alignée de manière indépendante. Si aucune des deux parties ne s'aligne, le read est considéré de qualité insuffisante. Si au moins une région s'aligne, il reste à déterminer le point de fusion. Un ensemble de fusions candidates est ainsi constitué. Notons que la méthode développée dans CRAC diffère largement de celle des autres outils. Contrairement à la tendance actuelle où les données *RNASeq* sont analysées par étapes successives, CRAC analyse les reads en une seule étape. Cette nouvelle approche intègre deux éléments, la position génomique et la couverture locale et utilise deux profils de k-mers pour détecter les mutations, *INDEL*, jonctions d'épissage et jonctions de fusion dans chaque lecture.

Plusieurs filtres sont réalisés afin d'éliminer un maximum de faux positifs. Les faux positifs surviennent essentiellement lorsqu'une lecture s'aligne à sa position d'origine et lorsque la deuxième lecture de la paire s'aligne erronément, en raison d'erreurs de séquençage ou de l'homologie des régions concernées. Les filtres classiquement réalisés sont la suppression de fusions au sein d'un même gène, entre des régions pas assez éloignées ou lues par un nombre insuffisant de lectures (la fusion doit être supportée par un certain nombre de lectures et des paires de lectures doivent s'aligner de part et d'autre de la jonction).

Troisième partie

Résultats

5 Altérations génétiques et épigénétiques dans la leucémie myélomonocytaire chronique et leur modulation par les agents déméthylants

Dans ce chapitre, nous décrivons les altérations génétiques que nous avons identifiées dans les monocytes de patients atteints de leucémie myélomonocytaire chronique. Nous rapportons ensuite une analyse de l'évolution de ces altérations au cours du temps ainsi que de l'influence des agents déméthylants sur la charge allélique des mutations, le niveau global de méthylation et l'expression des gènes dans les cellules étudiées. Les résultats sont présentés page 75 sous la forme d'un article en cours de révision dans le journal *Nature Communications*. Nous décrivons succinctement ces résultats dans les prochaines pages.

Nous avons séquencé les parties codantes du génome (N=49) et le génome entier (N=17) de cellules leucémiques et de cellules contrôles de patients *LMMC*. Les cellules leucémiques analysées étaient majoritairement des monocytes circulants, plus rarement des cellules mononuclées du sang et de la moelle. Les cellules contrôles utilisées étaient majoritairement des lymphocytes T, plus rarement des fibroblastes cutanés et des cellules de frottis de la muqueuse buccale. L'étude de ces échantillons a permis de déceler une moyenne de 14 mutations somatiques dans les régions codantes et 475 mutations somatiques dans les régions non répétées du génome des patients.

Les mutations synonymes des séquences codantes de l'*ADN* des cellules leucémiques n'ont pas été confirmées par reséquençage, mais simplement visualisées. L'une des raisons est que ces mutations ne sont le plus souvent pas récurrentes : seulement deux gènes portent des mutations silencieuses chez deux patients mais ils ne sont pas exprimés dans les monocytes ou $CD34^+$ de sujets sains ou patients *LMMC*. Les résultats suivants ne sont pas présentés dans l'article. Nous avons identifié 210 mutations synonymes chez les 49 patients, soit une moyenne de 4.3 (médiane : 4, étendue : 0-11) mutations par patient. Nous avons identifié des mutations dans des facteurs de transcription exprimés dans les monocytes ou $CD34^+$ de sujets sains ou patients *LMMC* : *GTF2E1*, *GTF2H1*, *SMARCAL1*, *HLX*, *MTERF* et *EIF4G1*. GTF2E1 est significativement sous-exprimé de plus de 3 fois dans les monocytes de patients *LMMC* comparés à des monocytes de sujets sains.

Les mutations validées dans les séquences codantes des gènes affectent des gènes régulateurs de l'épigénétique chez 92% des patients, des facteurs d'épissage chez 75% des patients et des gènes régulateurs de la signalisation cytokinique chez 59% des patients. En moyenne,

chaque patient a 3 gènes portant des mutations identifiées de manière récurrente dans la leucémie myélomonocytaire chronique. Nous avons de plus décelé des mutations récurrentes dans 9 gènes exprimés dans les monocytes ou les cellules plus immatures CD34+ : *PHF6*, *NF1*, *ETNK1*, *DOCK2*, *LUC7L2*, *ASXL2*, *ABCC9*, *HUWE1* et *TTN*. Cependant, les mutations du gène *TTN* ne sont probablement pas pertinentes : il s'agit d'un des plus longs gènes du génome, ses mutations sont donc plus fréquentes. Les parties codantes de ces huit gènes ont été séquencées dans une cohorte additionnelle de 180 patients *LMMC*. Nous avons trouvé les mutations avec les fréquences suivantes : *PHF6* (7.3%), *NF1* (6.1%), *ETNK1* (3.3%), *DOCK2* (2.1%), *ABCC9* (2.1%), *LUC7L2* (1.7%) et *HUWE1* (1.3%). Les mutations somatiques d'*ASXL2* n'ont pas été retrouvées dans la cohorte additionelle. Ces mutations ont été rapportées dans les leucémies aiguës (Micol et al. (2014)). Durant ma thèse, les mutations des gènes *ETNK1* (Gambacorti-Passerini et al. (2015)) et *LUC7L2* (Singh et al. (2013)) ont été identiées dans la leucémie myélomonocytaire chronique. *NF1* est un gène muté de manière récurrente dans les leucémies myélomonocytaires juvéniles, il est donc intéressant de le retrouver dans cette pathologie voisine.

Dans le génome, les mutations somatiques détectées sont essentiellement intergéniques (63.5%) et introniques (31.5%). Seulement 2.5% se trouvent dans les régions codantes. Les mutations somatiques détectées (N=8077) sont majoritairement des *SNV*. La signature mutationnelle des *SNV* (N=7568) a été analysée par Alexandrov. Sa méthode (Alexandrov et al. (2013b)) a intitialement permis l'identification de 30 signatures à travers une grande variété de génomes tumoraux (Alexandrov et al. (2013a)). L'application de cette méthode, combinée aux données répertoriées dans les génomes que son équipe a étudiés, a permis d'identifier trois processus à l'œuvre dans la leucémie myélomonocytaire chronique. Deux des processus ont été trouvés chez les 17 patients analysés et sont retrouvés fréquemment dans les cancers. Il s'agit des signatures 1 et 5 (Alexandrov et al. (2013a)), observées également dans plusieurs autres cancers et attribuées au vieillissement. Deux des 17 patients présentent une troisième signature (#31), qui n'avait jamais encore été identifiée. Le processus à l'origine de ces mutations n'est pas connu. Il est caractérisé par des mutations C :G>T :A en CpCpC et CpCpT et par un biais transcriptionnel important. Le ratio entre transitions (Ti) et transversions (Tv) $\frac{Ti}{Tv}$ de 1.96, très proche de celui des génomes non tumoraux, évoque encore une fois une signature du vieillissement.

Si, dans les parties codantes du génome, on observe une majorité de mutations non synonymes, on détecte aussi des pertes de codons stops et des insertions délétions. Les gains de codons stops sont rarissimes. Le ratio $\frac{Ti}{Tv}$ dans les parties codantes est de 2.7, ce qui est très proche de celui des exomes non tumoraux et renforce le rôle du vieillissement dans cette pathologie.

Nous nous sommes ensuite intéressés à l'évolution des mutations détectées dans cette première partie du travail. Nous avons étudié les séquences codantes du génome de 17 patients à plusieurs (2, 3 ou 4) temps. Les mutations des parties codantes du génome ne disparaissent pas au cours du temps, à l'exception de quelques rares sous-clones. Chez certains patients, de nouvelles lésions géniques apparaissent au cours du temps, qu'ils soient traités ou non traités par des agents déméthylants et, s'ils sont traités, qu'ils répondent ou non à leur traitement. Les mutations ne disparaissent donc pas chez les patients répondant à leur traitement et peuvent continuer à s'accumuler.

À partir de cette observation, nous avons décidé d'étendre l'étude de l'effet du traitement dans une cohorte de 9 patients que nous avons analysée à deux temps, par séquençage d'*ARN* polyadénylés et par *ERRBS*. Au premier temps, les 9 patients étaient non traités, au deuxième temps, 3 patients restaient non traités et les 6 autres étaient traités par agents déméthylants (Décitabine ou Azacitidine). Trois des patients étaient répondeurs et les trois autres étaient des patients stables, *i.e.* sans amélioration ni dégradation de l'état clinique, ce qui fait que les médecins continuaient le

traitement.

Nous avons d'abord étudié l'expression génique chez nos patients. Chez les patients répondeurs, nous avons identifié environ 500 gènes dont l'expression était dérégulée entre les 2 temps, contre une soixantaine seulement chez les patients stables et aucun chez les patients non traités. Nous avons ensuite étudié le niveau de méthylation chez ces mêmes patients. Chez les patients répondeurs, nous avons identifié environ 35000 régions méthylées de manière différentielle entre les deux temps, contre une centaine seulement chez les patients stables et 1 chez les patients non traités. Finalement, les traitements actuels ont un effet seulement épigénétique mais n'influent pas sur les mutations, ce qui pourrait expliquer les rechutes systématiques.

En conclusion, notre étude, qui avait été initiée pour découvrir de nouvelles mutations géniques récurrentes caractéristiques des leucémies myélomonocytaires chroniques a confirmé les résultats des études antérieures basées sur l'analyse de gènes candidats. Elle n'a pas identifié de mutation hautement récurrente. Elle a montré que l'évolution des altérations génétiques dans la leucémie myélomonocytaire chronique est relativement lente et peu influencée, semble-t-il, par les agents déméthylants. Ceux-ci semblent donc avoir un effet essentiellement épigénétique et non cytotoxique, n'éradiquant pas le clone leucémique mais restaurant une hématopoïèse équilibrée.

Mutation allele burden remains unchanged in chronic myelomonocytic leukemia responding to hypomethylating agents.

Jane Merlevede,[1,2*] Nathalie Droin,[1,2,3*] Tingting Qin,[4] Kristen Meldi,[4] Kenichi Yoshida,[5] Margot Morabito,[1,2] Emilie Chautard,[6] Didier Auboeuf,[7] Pierre Fenaux,[8] Thorsten Braun,[9] Raphael Itzykson,[8] Stéphane de Botton,[1,2] Bruno Quesnel,[10] Thérèse Commes,[11] Eric Jourdan,[12] William Vainchenker,[1,2] Olivier Bernard,[1,2] Noemie Pata-Merci,[3] Stéphanie Solier,[1,2] Velimir Gayevskiy,[13] Marcel E Dinger,[13] Mark J Cowley,[13] Dorothée Selimoglu-Buet,[1,2] Vincent Meyer,[14] François Artiguenave,[14] Jean-François Deleuze,[14] Claude Preudhomme,[10] Michael R Stratton,[15] Ludmil B Alexandrov,[15,16,17] Eric Padron,[18] Seishi Ogawa,[5] Serge Koscielny,[19] Maria Figueroa,[4] Eric Solary.[1,2,20]

* The two first authors contributed equally to this work

[1] INSERM U1170, Gustave Roussy, Villejuif, France.
[2] Gustave Roussy Cancer Center, Department of Hematology, Villejuif, France.
[3] INSERM US23, CNRS UMS3655, Gustave Roussy, Villejuif, France
[4] University of Michigan Medical School, Department of Pathology, Ann Arbor, Michigan.
[5] Department of Pathology and Tumor Biology, Kyoto University, Kyoto, Japan.
[6] Université Lyon 1, UMR CNRS 5558, Villeurbanne, France.
[7] Centre Léon Bérard, INSERM U1052, CNRS UMR5286, Lyon, France.
[8] Assistance Publique–Hôpitaux de Paris, Hôpital Saint-Louis, Paris, France.
[9] Assistance Publique-Hôpitaux de Paris, Hôpital Avicenne, Bobigny, France.
[10] Cancer Research Institute de Lille, INSERM U837, Lille, France.
[11] Institut de médecine régénératrice, Biothérapie et Institut de biologie computationnelle, INSERM U1040, Université de Montpellier, Montpellier, France;
[12] Centre Hospitalier Universitaire de Nîmes, Université Montpellier-Nîmes, Nîmes, France
[13] Kinghor Center for Clinical Genomics, Garvan Institute of Medical Research, Australia
[14] Centre National de Génotypage, Evry, France.
[15] Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, United Kingdom.
[16] Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, New Mexico.
[17] Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico.
[18] Malignant hematology, H. Lee Moffitt Cancer Center, Tampa FI, United States.
[19] Gustave Roussy Cancer Center, Department of Biostatistics, Villejuif, France.
[20] Faculty of Medicine, University Paris-Sud, Le Kremlin-Bicêtre, France.

Correspondence to:

Eric Solary, Inserm UMR 1170,

Gustave Roussy, 114, Rue Edouard Vaillant,

F-94805 Villejuif, France

Phone: +33.1.42.11.67.26

Fax : +33.1.42.11.52.40

email: eric.solary@gustaveroussy.fr

Keywords: Chronic myelomonocytic leukemia, gene mutations, gene expression, DNA methylation, hypomethylating agents.

The cytidine analogs azacytidine and 5-aza-2'-deoxycytidine (decitabine) are commonly used to treat myelodysplastic syndromes, with or without a myeloproliferative component. It remains unclear whether the response to these hypomethylating agents results from a cytotoxic or an epigenetic effect. This question is addressed in chronic myelomonocytic leukemia. A comprehensive analysis of the mutational landscape, combining whole exome and whole genome sequencing, identifies an average of 14 +/- 5 somatic mutations in coding sequences of sorted monocyte DNA and the signatures of three mutational processes. Serial sequencing demonstrates that the response to hypomethylating agents is associated with dramatic changes in DNA methylation and gene expression, without any decrease in the mutation allele burden, nor prevention of new genetic alteration occurence. Our findings indicate that cytosine analogs restore a balanced hematopoiesis without decreasing the size of the mutated clone, arguing for a predominantly epigenetic effect.

Introduction

Chronic myelomonocytic leukemia, a clonal hematopoietic malignancy that usually occurs in the elderly, is the most frequent myelodysplastic syndrome / myeloproliferative neoplasm [1]. Nonspecific cytogenetic abnormalities are observed in 30-40% of cases [2]. More than 30 candidate genes were identified to be recurrently mutated in leukemia cells [3-13]. Analysis of these recurrently mutated genes at the single cell level in 28 chronic myelomonocytic leukemia bone marrow samples identified the main features of the leukemic clone architecture, including the accumulation of mutations in the stem cell compartment with early clonal dominance, a low number of sub-clones, and a strong advantage to the most mutated cells with differentiation [4]. As in several other myeloid malignancies, *ASXL1* gene mutations demonstrated the strongest independent negative prognostic impact [14,15].

The median overall survival of chronic myelomonocytic leukemia patients is about 30 months, one third evolving to acute myeloid leukemia while the others die from the consequences of cytopenias. Allogeneic stem cell transplantation, which is the only curative therapy, is rarely feasible because of age. In patients ineligible for transplantation,

intensive chemotherapy results in low response rates and short response duration [16]. The cytidine analogs azacytidine and decitabine (5-aza-2'-deoxycytidine) were approved for the treatment of chronic myelomonocytic leukemia [17]. These azanucleosides were originally described as cytotoxic drugs, but low doses also cause DNA demethylation by inactivation of DNA methyltransferases [18,19]. It remains unclear whether the response to these drugs, which is always transient, results from a cytotoxic or an epigenetic effect.

To tackle this issue, we completed a comprehensive analysis of genetic alterations in chronic myelomonocytic leukemia cells by combining whole exome and whole genome sequencing. Then, we performed sequential whole exome and RNA sequencing together with DNA methylation analyses in untreated patients and patients treated with a hypomethylating drug. Clinical response to cytidine analogs was associated with a dramatic decrease in DNA demethylation, which was not observed when the disease remained stable on therapy. In responding patients, the size of the mutated clone remained unchanged, arguing for a predominantly epigenetic effect of these drugs.

Results

Comprehensive analysis of genetic alterations in coding regions

Since it remained uncertain whether the most frequent recurrent gene mutations had been all identified, we performed whole exome sequencing (WES) of paired tumor-control DNA from 49 chronic myelomonocytic leukemia cases (**Supplementary Fig. 1 and 2, Supplementary Table 1 and Table 2)** and validated 680 somatic mutations in 515 genes by deep re-sequencing (**Supplementary Table 3**). The average number of somatic mutations was 14+/-5 per patient (range: 4-23) (**Fig. 1a**). The most frequent alterations were somatic nonsynonymous single-nucleotide variants (N=515; 75.7%) (**Fig. 1b**). Most of the 618 variants were transitions (N=453, 73.3%) (**Fig. 1c**). We detected mutations affecting an epigenetic regulator gene in 45 out of 49 (91.8%) patients, a splicing machinery gene in 37 (75.5%) and a signal transduction gene in 28 (59.2%). Among the 36 genes found mutated in at least 2 patients, 19 had been previously identified in the context of chronic myelomonocytic leukemia, validating previous screens of mutations in candidate genes in this specific disease [3-13]. *TET2*, *SRSF2* and *ASXL1* were confirmed to be the most frequently mutated genes in chronic myelomonocytic leukemia [14].

Of the 17 other recurrently mutated genes, only 7 were actively transcribed in CD14-positive [20] and CD34-positive hematopoietic cells [according to Gene Expression Omnibus at http://www.ncbi.nlm.nih.gov/geo/]. These genes include *ABCC9 (*ATP-binding cassette, sub-family C member 9), *ASXL2 (*additional sex combs like 2), *DOCK2 (*dedicator of cytokinesis protein 2), *HUWE1 (*HECT, UBA and WWE domain containing 1, E3 ubiquitin protein ligase), *NF1 (*Neurofibromin 1), *PHF6 (*PHD finger protein 6), and *TTN* (Titin). Altogether, recurrent mutations were identified in 26 genes expressed in hematopoietic cells (**Fig.1d**). Constitutive truncating mutations in *TTN* gene were recently validated as a cause of dilated cardiomyopathy [21] and the variants identified in chronic myelomonocytic leukemia samples were validated by an independent method. Except this very large gene, the whole coding sequence of the 6 other genes, whose recurrent mutation in the context of chronic myelomonocytic leukemia had not been described previously, was deep sequenced in an additional cohort of 180 patients (**Supplementary Table 4**). Of the 229 studied patients, the most frequently mutated gene was *PHF6* (N= 17; 7.4%). *NF1* was altered in 14 (6.1%) patients. *DOCK2* and *ABCC9* mutations were detected respectively in 5 samples (2.1%), *HUWE1* mutations in 3 samples (1.3%) and *ASXL2* mutations in 2 samples (**Supplementary Table 5**). On average, each patient had 3.1 alterations (range: 1-7) among the 26 recurrently mutated genes identified in this series. Combinations are summarized in **supplementary Fig.3** and relationships with clinical and biological features in **supplementary Table 6**.

We extended this analysis by performing whole-genome sequencing (WGS) of paired tumor-control DNA from 17 patients. Of the 8077 somatic variants identified (**Fig. 2a**, **Supplementary Table 7** and **Table 8**), 207 were located in coding regions or splice sites (11.8 per patient) (**Fig. 2b**) and the combination of WES and WGS identified two additional recurrently mutated genes that are actively transcribed in hematopoietic cells, ten-eleven translocation 3 (*TET3)* and proline-rich coiled-coil 2B (*PRRC2B*). All these additional recurrent abnormalities may contribute to chronic myelomonocytic leukemia phenotype heterogeneity.

TET3 loss of function mutation

TET3 mutations are very infrequent in hematologic diseases [22,23] and were not detected in myeloid malignancies so far [24]. In the two patients with a mutated *TET3* gene, the two

alleles of *TET2* were also mutated. We further explored the functional consequences of $TET3^{R148H}$ identified in UPN22. Genetic analyses of $CD14^{+}$ cells at the single cell level (N=21) identified a complex repartition of *TET2* and *TET3* mutations, with $TET2^{S1708fs}$ being either alone or in combination with $TET3^{R148H}$, whereas $TET2^{L1819X}$ was detected in only one *TET3* wildtype cell (**Fig. 3a**). Expression of wildtype and $TET3^{R148H}$ alleles in HEK293T cells (**Fig. 3b**) demonstrated that $TET3^{R148H}$ mutation impaired the enzyme ability to promote 5-methylcytosine hydroxylation **(Fig. 3c).** Since many functional redundancies have been identified between TET2 and TET3 dioxygenases (for review see [25]), future studies are necessary to elucidate a potential cooperative interaction between TET2 and TET3 mutated alleles in diseased cells.

Comprehensive analysis of genetic alterations in non-coding regions.

Further analysis of WGS data indicated that, on average, chronic myelomonocytic leukemia cells carried 475 (range: 27-854) somatic variants in their DNA (**Fig. 2a**), 6.3% being short insertions and deletions. These variants were mostly in intergenic (63.5%) and intronic (31.5%) regions (**Fig. 2b**). Somatic single nucleotide variants (93.7%) were mostly transitions (66.3%) (**Fig. 2c**), and synonymous base changes represented 24.1% of the identified variants (**Fig. 2d**). Our computational framework for extracting mutational signatures [26] identified the signatures of three mutational processes (**Fig. 2e**). Two (signatures 1 and 5) were previously observed [27] and believed to be due to clock-like mutational processes operative in normal somatic tissues. Interestingly, we identified in 2 cases a novel mutational signature (signature 31) characterized by C:G>T:A mutations at CpCpC and CpCpT (mutated based underlined) and exhibiting a strong transcriptional strand bias (**Supplementary Fig. 4**). We did not detect any recurrent alteration in non-coding regions, as described in other tumor types [28,29,30]. We identified 21 potential hotspot regions with at least 2 variants in distinct samples being at most 250 bp far (**Fig. 2f**). Nine were in the coding sequence of recurrently mutated genes, and 3 in non-coding regions of genes transcribed in hematopoietic cells (*PDS5A*, *ZFP36L2, NHLRC2*). Finally, we detected 147 variants in promoters and 37 variants in permissive enhancers, of which three showed activity in blood cells (**Supplementary Table 9**) [31].

Serial whole exome analyses

WES of sorted monocyte DNA was repeated in 17 patients. The mean time between two analyses was 14+/-8 months (range 4-32). Six patients received supportive care, whereas 11 were treated with either azacytidine (N=5) or decitabine (N=6). The number of serial analyses per patient ranged from 2 to 5 (**Supplementary Fig. 1** and **Supplementary Table 10**). The mean duration of treatment was 21±13 months (range 5-47). One or two WES were performed before treatment, subsequent analyses being performed on therapy in samples collected immediately before the next cycle. Five of the treated patients demonstrated a response at the time of sampling ("responders"), including 1 complete response (UPN32), 3 marrow complete responses with hematological improvement, and 1 marrow complete response without hematological improvement (UPN34). In the 6 other patients, the disease remained stable on therapy, without hematological improvement ("non-responders") [19,32]. In total, we performed 27 serial WES analyses. In 17 cases, we did not detect any change in gene mutations as compared to the previous analysis, the mutated allele burden remaining stable in all patients but two (UPN23 and UPN47) (**Fig. 4**). In responding patients, hypomethylating agents did not decrease the mutated allele burden in circulating monocytes. In 8 cases, we detected changes in the number of mutated genes, including 3 untreated, 3 non-responders with a stable disease and one responder (**Fig. 4** and **Supplementary Fig. 5-21**). The latter was a 74-year old man (UPN34) with 12 somatic mutations at diagnosis who successively acquired mutations in *CNTN4* and *RAD21* genes, then in *KRAS*, *CNTN6*, *PCDHGA6* genes while being in complete marrow response without hematological improvement. The last exome analysis, performed in acute transformation, identified an *EZH2/ETV6* mutated subclone (**Supplementary Fig. 21**). UPN46 was analyzed first while being untreated, showing the disappearance of a subclone with *ARID2* and *NRAS* mutations while another clone with *NRAS, ROBO2, FAT1* and *SGSM2* mutations expanded. This patient was subsequently treated with decitabine and responded to treatment, without change in mutation number and allele burden (**Fig. 4c, Supplementary Fig. 18 and Fig. 22**).

In one additional patient who demonstrated a long and complete response to AZA, then progressed to AML (described in more details in the methods section), serial WGS of bone marrow mononucleated cells [33] was performed. Prior to AZA therapy, somatic variants in *TET2*, *EZH2* and *CBL* genes were identified. In a best response sample, a striking

stability of variant allele frequency was observed. At the time of progression, a loss of heterozygocity of mutated *EZH2* was detected, together with the acquisition of a mutation in *ASXL1*, and a whole loss of chromosome 7, which was confirmed by serial cytogenetic analysis (**Fig. 5**). This observation emphasizes the lack of genetic response to AZA and the possibility to detect genetic progression on therapy, preceding progression to acute leukemia.

Gene expression and DNA methylation

In 9 of these patients, we performed serial RNA sequencing **(Fig. 6, Supplementary Table 11** and **Table 12)**, the first sample being collected before treatment. Three remained untreated, and six were treated with a hypomethylating drug, the second sample being collected on therapy. Of the six treated patients, three were responders, the three others remaining on therapy with stable disease (non-responders). We measured the effect of time on gene expression. We noticed a strong impact of treatment in responders, with 513 differentially expressed genes, whereas only 63 genes were differentially expressed in treated patients with stable disease (non-responders), and none in untreated patients (**Table 1, Fig. 6a, 6b** and **Supplementary Table 13).** The proportions of significantly differentially expressed genes between the groups were all significantly different ($P<10^{-10}$, Chi-squared test). RT-QPCR analysis validated all the tested up-regulated genes in an extended cohort of 6 responders compared to 10 patients with stable disease (**Fig. 6c & Supplementary Fig. 23 and Fig. 24**)

Finally, we explored the effect of time on methylation status in the same samples by using the enhanced reduced representation bisulfite sequencing assay (**Fig. 7**). Differentially methylated regions (DMRs) between the two time points were defined by a more than 25% change in methylation and a False Discovery Rate ≤ 10%. Differential methylation was detected almost exclusively in the three responding patients (**Fig. 7b, 7d,** and **7e**). The number of DMRs remained low in non-responding patients with a stable disease under therapy (**Fig. 7a, 7c** and **7e**) and no change was identified in untreated patients (**Table 1, Supplementary Fig. 25** and **Supplementary Table 14**). Changes observed in responding patients were predominantly demethylation, whereas changes detected in treated patients with a stable disease included both gains and losses of DNA methylation (**Supplementary Fig. 25).** In responders, DMRs were significantly depleted in promoters and in CpG islands

while being enriched in generic enhancers (**Supplementary Fig. 26**). Some overlap was detected between differentially methylated regions and changes in gene expression in responders, which was not observed in non-responders (**Fig. 8**).

Discussion

This first comprehensive analysis of genetic alterations in chronic myelomonocytic leukemia cells demonstrates that azanucleosides, although inducing dramatic changes in DNA methylation and gene expression in responding patients, do not reduce the mutated allele burden, nor permit the re-expansion of wild-type hematopoietic cells.

Previous screening of candidate genes identified somatic mutations in *TET2, ASXL1,* and *SRSF2* genes as the most frequent recurrent events in chronic myelomonocytic leukemia cells [4]. Our comprehensive analysis validates this molecular fingerprint and identifies additional recurrent abnormalities that may contribute to the disease phenotype heterogeneity. Several of the most recurrent mutations identified in leukemic cells were associated with age-related clonal hematopoiesis [34-36] or “silent” pre-leukemic clones [37-39]. The bias in myeloid differentiation towards the granulomonocytic lineage that characterizes chronic myelomonocytic leukemia could be related to the expansion of such a clone, *e.g.* due to early clonal dominance of *TET2* [4,40]. In this setting, the occurrence of an additional mutation resulting in a stringent arrest of differentiation leads to acute-phase disease [39,41], as illustrated by sequential analyses in UPN34 who partially responded to decitabine for 2 years until the emergence of an *EZH2* / *ETV6* mutated subclone and an acute leukemia phenotype. Importantly, this observation indicates that the response to a hypomethylating agent does not prevent the accumulation of genetic damage in the leukemic clone.

The number of genetic alterations identified in the genome of chronic myelomonocytic leukemia cells was close to that observed in other hematological malignancies [27]. Most somatic variants identified were transitions, with a predominance of C:G->T:A, and a mutational signature suggesting that the historical mutational process was related mostly to ageing [27]. Accordingly, the number of variants identified in juvenile chronic myelomonocytic leukemia, another myeloproliferative neoplasm/myelodysplastic disease that occurs in young children, is much lower than that measured in chronic

myelomonocytic leukemia [42].

Although these results do not exclude some cytotoxic effect of azanucleosides, their epigenetic activity appears to play a central role in restoring a more balanced hematopoeisis in the 30-40% of chronic myelomonocytic leukemia patients who respond to these drugs [18,19]. Immunophenotyping analyses already suggested that these drugs could eliminate bulk blast cells without eradicating leukemia stem and progenitor cells in AML patients [43] and did not correct $CD34^+$ cell immunophenotypic aberrancies in chronic myelomonocytic leukemia patients [44]. Mutations in epigenetic genes observed in almost every chronic myelomonocytic leukemia case lead to DNA hypermethylation [45] and epigenetically controlled changes in gene expression contribute to the disease phenotype, as demonstrated for *transcription intermediary factor 1γ* (*TIF1γ*) gene whose epigenetic down-regulation was identified in a fraction of patients, and whose deletion in the myeloid compartment induces a chronic myelomonocytic leukemia phenotype in the mouse [46]. Clinical response to hypomethylating drugs is associated with a re-expression of this gene when initially down-regulated [46], indicating that hypomethylating drugs can suppress epigenetic changes that contribute to the disease phenotype. This epigenetic effect could decrease the competitiveness of the most mutated cells in the progenitor and stem cell compartment [4,41] but not the mutated allele burden in the mature cell compartment.

Clinical trials have shown that 30 to 40% of chronic myelomonocytic leukemia patients respond to azanucleosides [16,19]. Since epigenetic changes were observed only in responders, specific patterns of epigenetic changes may be amenable to reversion by azanucleosides [18]. We have shown that differentially methylated non-promoter regions of DNA at baseline distinguished responders from non-responders to decitabine [47], whereas the pattern of somatic mutations did not [19]. Some epigenetic patterns could also prevent the activity of hypomethylating drugs by either decreasing the expression of human nucleoside transporters and metabolic enzymes needed for their activation such as cytidine and deoxycytidine kinases and cytidine deaminase [17,48] or increasing the expression of genes encoding cytokines such as CXCL4 and CXCL7 that, when released, could antagonize the drug effects [47]. In two responding patients, prolonged administration of azanucleosides, although improving hematopoiesis, did not prevent the accumulation of genetic events, ultimately leading to acute transformation, indicating that these drugs do not prevent genetic evolution of the leukemic clone. Further analyses are

needed to determine if they could even promote such genetic evolution.

The present findings have clinical implications. First, prolonged administration of hypomethylating drugs may not have any benefit in chronic myelomonocytic leukemia patients when haematological improvement is not observed after a few cycles. Secondly, these drugs could increase the survival of responding patients by restoring a more balanced hematopoiesis, but they might not prevent the occurrence of new genetic events leading to acute transformation. Finally, better analysis of how these drugs modulate the immunogenicity of mutated cells could lead to combination of hypomethylating agents with immune checkpoint blockers as nucleoside analogs render the cells more immunogenic through inducing the expression of cancer testis antigens [49], promoting the demethylation of programmed death-1 (PD-1) immune checkpoint molecule [50], and inducing retrovirus activation [51,52], suggesting that an interaction of epigenetic drugs and immunotherapeutic approaches [53] might be considered. Our results also raise the question on whether epigenetic targeting molecules currently developed to treat haematological malignancies [54,55] will eradicate mutated cells or erase the epigenetic consequences of these mutations, leading to the transient restoration of a more balanced hematopoiesis.

Methods

Patients. Peripheral blood and bone marrow samples were collected on ethylenediaminetetraacetic acid from 245 patients with a chronic myelomonocytic leukemia (CMML) diagnosis according to the World Health Organisation criteria [1]. When indicated, several peripheral blood samples were collected sequentially from a given patient (**Supplementary Fig. 1**). We initially performed whole exome sequencing (WES) in 49, whole genome sequencing (WGS) in 17, and validation of recurrent mutations by deep sequencing in 180 cases. Serial WES were performed in 17 patients, including 6 untreated and 11 treated with either decitabine (N = 6; EudraCT 2008-000470-21 GFM trial; NCT01098084; https://www.clinicaltrials.gouv/) [19] or azacytidine (N = 5; following the European Medicines Agency approval; EMEA/H/C/000978). Responses were classified according to the International Working Group 2006 criteria [32]. Patients with stable disease without hematological improvement remained treated until progression [18]. When indicated, sequential RNA sequencing and DNA methylation analysis [47] were performed. In treated patients, samples were collected immediately before the following drug cycle. All the procedures were approved by the relevant ethics committees, and written informed consent was obtained from each patient. Data collected from French and Japanese patients were analyzed homogeneously. Patient characteristics are in **Supplementary Table 1,** the flow chart of analyses in **Supplementary Fig. 1.**

Cell sorting. Bone marrow (N=9) or peripheral blood (N=7) mononucleated cells were separated on Fycoll-Hypaque. Peripheral blood $CD14^+$ monocytes were sorted with magnetic beads and the AutoMacs system (Miltenyi Biotech, Bergish Gladbach, Germany) [46]. Control samples were peripheral blood CD3-positive T lymphocytes sorted with the AutoMacs system or buccal mucosa cells (N=3) or skin fibroblasts (N=12). All the samples used in the validation cohort (N=180) were sorted peripheral blood $CD14^+$ monocytes. DNA and RNA were extracted from cell samples using commercial kits. Sorted monocytes were chosen for DNA sequencing on the basis of our previous analysis of chronic myelomonocytic leukemia clonal architecture showing the grow advantage to the most mutated cells [4] and flow cytometry analysis of peripheral blood monocytes showing limited phenotypic alteration in the classical monocyte population in patients treated with

hypomethylating drugs, even though responders have more intermediate and non-classical monocytes [20]. In one patient, bone marrow mononucleated cells were used for serial whole genome sequencing. *TET2* and *TET3* gene sequencing in UPN22 were performed in single $CD14^+$ cells sorted using C1 (Fluidigm) after whole genomic DNA amplification.

Functional analysis of mutated *TET3*. pcDNA3.1-*TET3R1548H* was generated using Q5® site-directed mutageneis (New England Biolabs Evry, France) before transfecting HEK293T cells with constructs encoding wildtype or mutated *TET3*. After 2 days in culture, DNA was extracted and 5 hydroxymethylcytosine was detected as previously described [40].

Whole exome sequencing (WES). We performed WES in 49 patients at diagnosis. In 17 of them, 2 to 4 serial analyses were done. One µg of genomic DNA was sheared with the Covaris S2 system (LGC Genomics, Molsheim, France). DNA fragments were end-repaired, extended with an 'A' base on the 3′ end, ligated with paired-end adaptors, and amplified (six cycles) using a Bravo automated platform (Agilent technologies). Exome-containing adaptor-ligated libraries were hybridized for 24 hours with biotinylated oligo RNA baits, and enriched with streptavidin-conjugated magnetic beads using SureSelect (Agilent technologies, Les Ulis, France). The final libraries were indexed, pooled and paired-ends (2x100bp) sequenced on Illumina HiSeq-2000 (San Diego, CA). In 9 cases, WES was performed in Japan following a previously described protocol [56]. The mean coverage in the targeted regions was 112x (**Supplementary Table 2**). Two individual cases have been reported in our previous studies [4,8].

Whole exome sequencing (WES) analysis. Raw reads were aligned to the reference human genome hg19 (Genome Reference Consortium GRCh37) using BWA 0.5.9 (Burrows-Wheeler Aligner) backtrack algorithm with default parameters. Polymerase chain reaction (PCR) duplicates were removed with Picard (http://picard.sourceforge.net) version 1.76. Local realignment around indels and base quality score recalibration were performed using GATK 2.0.39 (Genome Analysis ToolKit). Statistics on alignment and coverage are given in **Supplementary Table 2**. Single nucleotide variants (SNVs) and indels were

called with VarScan2 somatic 2.3.2 [57]. Reads and bases with a Phred-based quality score ≤20 were ignored. Variants with somatic p-value below 10^{-4} (or 10^{-3} for samples with mean coverage <100x or contamination >15% in $CD3^+$ control sample) were reported. In addition to the Fisher Exact test of VarScan, we required (variant allele frequency in the tumor sample - variant allele frequency in the normal sample) ≥15% to distinguish somatic from germline variations. Variants were annotated with Annovar. Mutations were searched in 1000G (April 2012) and Exome Sequencing Project (ESP5400). Conservation of the position was predicted by PhyloP and the effect of the mutation was predicted by SIFT, Polyphen2, LRT and MutationTaster. We excluded variants reported in dbSNP version 129, filtered variants located in intergenic, intronic, untranslated regions and non-coding RNA regions, and removed synonymous SNVs and variants with mapping ambiguities. A mutation was reported as present if VAF≥4%.

Targeted deep sequencing. Regarding exome validation, Ion AmpliSeq™ Custom Panel Primer Pools were used to perform multiplex PCR for preparation of amplicon libraries. Briefly, 20 ng of DNA per primer pool quantified using a Qubit Fluorometer (Invitrogen, Carlsbad, CA) were used in the multiplex PCR. Unique indexed libraries per sample were generated, quantified by Qubit, pooled, and run on an Ion 318™ Chip using the Ion PGM™ Sequencer (Life Technologies). Seventy one percent of the candidates for somatic mutation were confirmed by deep resequencing at a mean coverage of 759x. In total, we validated 680 somatic mutations (**Supplementary Table 3**). Also, the whole coding regions of genes found mutated in at least 2 patients and expressed in myeloid cells were deep sequenced (mean coverage, 690x) in a cohort of 180 chronic myelomonocytic leukemia patients (**Supplementary Table 4** and **Supplementary Table 5)**. Ion AmpliSeq™ Custom Panel Primer Pools were used (10 ng of gDNA per primer pool) to perform multiplex PCR. Libraries were generated with addition of paired-end adaptors (NEXTflex, Bioo Scientific) before paired-end sequencing (2x150 bp reads) using an Illumina MiSeq flow cell and the onboard cluster method (Illumina, San Diego, CA). Quality of reads was evaluated using FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/). Raw reads were filtered with Trimommatic 0.30 [58] to remove adaptors, truncate any read whose average quality on a sliding window (6 bases) was ≤20, remove the start and the end of a read if ≤20 and any

read with an average quality ≤20 or a length<36. Statistics on alignment and coverage are given in **Supplementary Table 2** and detailed analysis of each studied variant in **Supplementary Table 3**. Targeted resequencing was analysed similarly to WES except the suppression of PCR duplicates. We added the following public databases: ESP 6500, dbSNP 138, COSMIC 68 (Catalogue Of Somatic Mutations In Cancer) and ClinVar (20140303).

Prediction of driver genes. We applied DrGaP (driver genes and pathways) [59] to synonymous and nonsynonymous somatic variants (889 in 694 genes) to measure the probability of each variant to occur by chance. Among the 22 genes with FDR≤10%, 20 were mutated in at least 2 patients and actively transcribed in myeloid cells (*MIER* and *FIBIN* genes carried 2 variants in a unique patient, respectively). Six of the 26 recurrently mutated and actively transcribed genes were mutated in only 2 patients: *SH2B3* (FDR=0.22), *PHF6* (FDR=0.22), *DOCK2* (FDR=0.30), *ABCC9* (FDR=0.36), *HUWE1* (FDR=0.78) and *TTN* (FDR=0.88).

Whole genome sequencing (WGS). We performed WGS in 17 patients at diagnosis, including one already studied by WES. Genomic DNA (1µg) was sheared to 300-600 bp (average size= 398±14bp) using a Covaris E210 (Covaris, Woburn, Massachusetts, USA). Libraries for 101 bp paired-end sequencing were prepared according to the Truseq PCR free protocol (Illumina). Library quality was evaluated by qPCR for quantification (Kapa Biosystems Ltd. London, UK) and by low output sequencing on Miseq (Illumina) for clusterisation efficiency. Samples were loaded on HiSeq 2000 and sequenced. Quality of reads was evaluated using FastQC. Sequences were filtered with Trimommatic. Reads were aligned to the reference genome hg19 using BWA mem algorithm 0.7.5a with default parameters. The PCR duplicates were removed with Picard 1.94. Local realignment and base quality score recalibration were performed using GATK 2.7.4. The mean coverage of all the samples was 31x. Detailed statistics on alignment and coverage are given in **Supplementary Table 7**. Somatic SNVs were identified by SomaticSniper 1.0.3 [60], VarScan2 2.3.7 and Strelka 1.0.14 [61]. We conserved somatic variants with ≥15x in normal, ≥6x in tumor, and ≥3 reads supporting the variant. We used a SomaticScore Tumor ≥30 for SomaticSniper, a Somaticpvalue≤0.01 for VarScan2 and a

QSS_N/QSI_NT≥15 for Strelka. We ran Strelka with the following parameters: ssnvNoise = 0.000000005, sindelNoise = 0.00000001, ssnvPrior = 0.001, sindelPrior = 0.001 and extraStrelkaArguments: -used-allele-count-min-qscore 20 and min-qscore 20. The other parameters were set by default. We removed SNVs annotated as SpanDel, BCNoise or DP in FILTER field. We excluded INDELs reported as OVERLAP or defined as Repeat, iHpol, BCNoise or DP in the FILTER field. In addition, we required (variant allele frequency in the tumor sample - variant allele frequency in the normal sample) ≥20% and excluded the variants whose allele frequency in the normal sample was ≥15% to differentiate somatic from germline mutations. We removed the variants located in low complexity regions, immunoglobulin loci [as reported on http://www.genecards.org/ for *TCRA, TCRB, TCRG, IGH, IGL, IGK*] and genes in which false positives have been frequently detected by new generation sequencing [62]. By removing low complexity regions, as defined in the masked genome chromOut.tar.gz generated by repeatMasker (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/), we removed 45% of the genome, thereby eliminating 74%, 69% and 71% of the SNVs detected by using SomaticSniper, VarScan2 and Strelka, respectively, and 87% and 77% of the indels identified by VarScan2 and Strelka, respectively. In subsequent analyses, the single nucleotide variants and indels identified by combining these stringent algorithms were used.

Potential hotspots, promoters and enhancers. First, sequential windows were used to calculate the probability for a 250bp region to carry at least two variants in two distinct patients among 17 patients. The probability to find at least 2 mutations in one of the $6.82.10^6$ windows of 250 bp defined in non-repeated regions of the genome among 17 patients was 10^{-3}. Secondly, we defined a potential hotspot region as a region in which, in a sequence shorter than 250bp, 2 variants were identified in at least two patients. With that method, from the 8,077 variants detected (**Supplementary Table 8)**, we identified 21 clusters of variants on a same chromosome at most 250 bp far, defining potential hotspots (**Supplementary Table 9**). We detected 147 variants in 144 distinct promoter regions (from -2000 bp before to + 200 bp after the translation starting site obtained from UCSC website on 11/07/2014) and 37 variants in the 43,011 enhancer regions reported in [31], of which three were located in the 3,795 enhancers whose activity is >=5% in blood

and >=5% in monocytes.

Independent case report from Lee Moffitt Cancer Center: A 57-year female patient progressed approximately 10 months after diagnosis of a type-1 chronic myelomonocytic leukemia according to the WHO definition with normal cytogenetics, prompting the initiation of 5-azacitidine therapy. After 4 cycles of therapy, the patient had a complete remission that persisted for 30 cycles. Disease progression was suspected because of a declining platelet count and confirmed by an increase in bone marrow myeloblasts. 5-azacitidine was discontinued and the patient transformed to AML 8 months later. Bone marrow monucleated cells [33] were collected before the treatment start, during complete response and at progression. WGS was performed on five lanes for each leukemia sample, and two lanes for the CD3+ germline on the Illumina HiSeq X platform, for a total 22 lanes. The goal was to achieve 125x depth, and 60x, respectively. Sequencing data was aligned to b37d5 reference genome with BWA MEM, and duplicates were marked, and multiple lanes merged using novosort. Somatic SNV and INDEL variant calling was performed using Strelka for tumor normal pairs. Somatic copy number variants, loss of heterozygosity regions, ploidy and purity were determined using Sequenza. Freebayes with minimum VAF=0.01 was used to generate variants from individual samples, and to assess the number of clones. Variants were annotated using Variant Effect Predictor. Phylosub was used to reconstruct the evolutionary lineage of samples, using either high, or medium- and high-impact variants (loss of function, vs missense, respectively).

RNA Sequencing (RNA-Seq). Sequential RNA-seq was performed on 18 samples (9 patients) with high quality RNA (RNA Integrity Score ≥ 7.0 as determined by the Agilent 2100 Bioanalyzer). RNA was quantified using a Qubit Fluorometer (Invitrogen, Cergy-Pontoise, France). RNA-Seq libraries were prepared using the SureSelect Automated Strand Specific RNA Library Preparation Kit per manufacturer's instructions (Agilent technologies) and a Bravo automated platform (Houston, TX). Briefly, 150 ng of total RNA sample was used for poly-A mRNA selection using oligo(dT) beads and subjected to thermal mRNA fragmentation. The fragmented mRNA samples were subjected to cDNA synthesis and further converted into double stranded DNA that was used for library preparation. The final libraries were bar-coded, purified, pooled together in equal

concentrations, and subjected to paired-end (101bp) sequencing on HiSeq2000 (San Diego, CA). Two separate samples were multiplexed into each lane. Quality of reads was evaluated using FastQC.

RNA-Seq analysis. Sequences were filtered with Trimommatic and alignment was performed with Tophat2 version 2.0.9 [63] and Bowtie2 version 2.1.0 [64]. The filtered reads were aligned to a reference transcriptome (downloaded from UCSC website on 12/20/2013). The remaining reads were split and segments were aligned on the reference genome, as described [63]. In average, 88.95% of reads were aligned (**Supplementary Table 11**) and counted with HTSeq (v0.5.4p5) [65] using the following parameters: --mode=intersection-nonempty --minaqual=20 –s. Differential expression analysis was performed using DESeq2 package version 1.6.3 [66] with R statistical software version 3.1.2. To study the effect of time in each of the 3 groups (**Supplementary Table 13**), we used a generalized linear model to explain the counting Y_i: Y_i ~ Group:Patient+Time+Group+Group:Time where Group indicates the status (untreated, responders, stable disease). We used independent filtering to set aside genes that have no or little chance to be detected as differentially expressed. To test the effect of time in each group, we used 3 contrasts defined as linear combinations of factor level means. Validation of RNA-Seq data was performed by qPCR analysis in a selection of 8 genes, using 3 independent genes as reporters (**Supplementary Fig. 23**).

Genome-wide DNA methylation by ERRBS Twenty-five nanograms of high-molecular weight genomic DNA were used to perform the ERRBS assay as previously described [67] and sequenced on a HiSeq2000 Illumina sequencer. 50 bp reads were aligned against a bisulfite-converted human genome (hg19) using Bowtie and Bismark [68]. Downstream analysis was performed using R version 3.0.3, Bioconductor 2.13 and the MethylSig 0.1.3 package. Only genomic regions with coverage between 10X and 500X were used for the downstream analysis (**Supplementary Table 14**). Differentially methylated regions (DMR) were identified by first summarizing the methylation status of genomic regions into 25-bp tiles and then identifying regions with absolute methylation difference ≥25% and false discovery rate (FDR) <10%. DMRs were annotated to the RefSeq genes using the following criteria: (i) DMRs overlapping with a gene were annotated to that gene, (ii)

intergenic DMRs were annotated to all neighboring genes within a 50-kb window, and (iii) if no gene was detected within a 50-kb window, then the DMR was annotated to the nearest TSS.

References

1. Vardiman, J.W. *et al*. The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood* **114**, 937-951(2009).
2. Padron, E. & Steensma, D.P. Cutting the cord from myelodysplastic syndromes: chronic myelomonocytic leukemia-specific biology and management strategies. *Curr. Opin. Hematol*. **22**,163-170 (2015).
3. Jankowska, A.M. *et al*. Mutational spectrum analysis of chronic myelomonocytic leukemia includes genes associated with epigenetic regulation: UTX, EZH2, and DNMT3A. *Blood* **118**, 3932-3941 (2011).
4. Itzykson, R. *et al*. Clonal architecture of chronic myelomonocytic leukemias. *Blood* **121**, 2186-2198, (2013).
5. Kon, A. *et al*. Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. *Nat. Genet.* **45**, 1232-1237 (2013).
6. Yoshida, K. *et al*. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**, 64-69 (2011).
7. Gelsi-Boyer, V. *et al*. Genome profiling of chronic myelomonocytic leukemia: frequent alterations of RAS and RUNX1 genes. *BMC Cancer* **8**, 299 (2008).
8. Kosmider, O. *et al*. Mutation of the colony-stimulating factor-3 receptor gene is a rare event with poor prognosis in chronic myelomonocytic leukemia. *Leukemia* **27**, 1946-1949 *(*2013).
9. Dunbar, A.J. *et al*. 250K single nucleotide polymorphism array karyotyping identifies acquired uniparental disomy and homozygous mutations, including novel missense

substitutions of c-Cbl, in myeloid malignancies. *Cancer Res.* **68**, 10349-10357 (2008).

10. Gómez-Seguí, I. *et al*. Novel recurrent mutations in the RAS-like GTP-binding gene RIT1 in myeloid malignancies. *Leukemia* **27**, 1943-1946 (2013).
11. Klinakis, A. *et al*. A novel tumour-suppressor function for the Notch pathway in myeloid leukaemia. *Nature* **473**, 230-233 (2011).
12. Singh, H. *et al*. Putative RNA-splicing gene LUC7L2 on 7q34 represents a candidate gene in pathogenesis of myeloid malignancies. *Blood Cancer J*. **3**, e117 (2013).
13. Gambacorti-Passerini, C.B. *et al*. Recurrent ETNK1 mutations in atypical chronic myeloid leukemia. *Blood* **125**, 499-503 (2015).
14. Itzykson, R. *et al*. Prognostic score including gene mutations in chronic myelomonocytic leukemia. *J. Clin. Oncol.* **31**, 2428-2436 (2013).
15. Patnaik, M.M. *et al.* ASXL1 and SETBP1 mutations and their prognostic contribution in chronic myelomonocytic leukemia: a two-center study of 466 patients. *Leukemia* **28**, 2206-2212 (2014).
16. Padron, E. & Steensma, D.P. Cutting the cord from myelodysplastic syndromes: chronic myelomonocytic leukemia-specific biology and management strategies. *Curr. Opin. Hematol.* **22**, 163-170 (2015).
17. Navada, S.C., Steinmann, J., Lübbert, M. & Silverman, L.R. Clinical development of demethylating agents in hematology. *J. Clin. Invest.* **124**, 40-46 (2014).
18. Tsai, H.C. *et al*. Transient low doses of DNA-demethylating agents exert durable antitumor effects on hematological and epithelial tumor cells. *Cancer Cell* **21**, 430-446 (2012).

19. Braun, T. *et al*. Molecular predictors of response to decitabine in advanced chronic myelomonocytic leukemia: a phase 2 trial. *Blood* **118**, 3824-3831 (2011).

20. Selimoglu-Buet, D. *et al*. Characteristic repartition of monocyte subsets as a diagnostic signature of chronic myelomonocytic leukemia. *Blood* **125**, 3618-3626 (2015).

21. Hinson, J. T. *et al*. Titin mutations in iPS cells define sarcomere insufficiency as a cause of dilated cardiomyopathy. *Science* **349**, 982-986 (2015).

22. Quesada, V. *et al*. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat. Genet*. **44**, 47-52 (2011).

23. Palomero, T. *et al*. Recurrent mutations in epigenetic regulators, RHOA and FYN kinase in peripheral T cell lymphomas. *Nat. Genet.* **46**, 166-170 (2014).

24. Abdel-Wahab, O. *et al*. Genetic characterization of TET1, TET2, and TET3 alterations in myeloid malignancies. *Blood* **114**, 144-147 (2009).

25. Ko, M. *et al*. TET proteins and 5-methylcytosine oxidation in hematological cancers. *Immunol. Rev.* **263**, 6-21 (2015).

26. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J. & Stratton, M.R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep*. **3**, 246-259 (2013).

27. Alexandrov, L.B. *et al*. Signatures of mutational processes in human cancer. *Nature* **500**, 415-421 (2013).

28. Horn, S. *et al*. TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959-961 (2013).

29. Schulze, K. *et al*. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat. Genet.* **47**, 505-511 (2015).

30. Melton, C., Reuter, J.A., Spacek, D. V. & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat. Genet.* **47**, 710-716 (2015).

31. Andersson, R. *et al*. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461 (2014).

32. Cheson, B.D. *et al*. Clinical application and proposal for modification of the International Working Group (IWG) response criteria in myelodysplasia. *Blood* **108**, 419-425 (2006).

33. Mohamedali, A. M. *et al*. High concordance of genomic and cytogenetic aberrations between peripheral blood and bone marrow in myelodysplastic syndrome (MDS). *Leukemia* **29**, 1928-1938 (2015).

34. Jaiswal, S. *et al*. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488-2498 (2014).

35. Genovese, G. *et al*. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477-2487 (2014).

36. Xie, M. *et al*. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med*. **20**, 1472-1478 (2014).

37. Jan, M. *et al*. Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Sci. Transl. Med*. **4**, 149ra118 (2012).

38. Shlush, L.I. *et al*. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **506**, 328-333 (2014).

39. Potter, N.E. & Greaves, M. Cancer: Persistence of leukaemic ancestors. *Nature* **506**, 300-301 (2014).

40. Quivoron, C., *et al*. TET2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. *Cancer Cell* **20**, 25-38 (2011).

41. Itzykson, R. & Solary, E. An evolutionary perspective on chronic myelomonocytic leukemia. *Leukemia* **27**, 1441-1450 (2013).

42. Sakaguchi, H. *et al*. Exome sequencing identifies secondary mutations of SETBP1 and JAK3 in juvenile myelomonocytic leukemia. *Nat. Genet.* **45**, 937-941 (2013).

43. Craddock, C. *et al*. Azacitidine fails to eradicate leukemic stem/progenitor cell populations in patients with acute myeloid leukemia and myelodysplasia. *Leukemia* **27**, 1028-1036 (2013).

44. Shen, Q. *et al*. Flow cytometry immunophenotypic findings in chronic myelomonocytic leukemia and its utility in monitoring treatment response. *Eur. J. Haematol.* **95**, 168-176 (2015)

45. Figueroa, M.E. *et al*. MDS and secondary AML display unique patterns and abundance of aberrant DNA methylation. *Blood* **114**, 3448-3458 (2009).

46. Aucagne, R. *et al*. Transcription intermediary factor 1γ is a tumor suppressor in mouse and human chronic myelomonocytic leukemia. *J. Clin. Invest.* **121**, 2361-2370 (2011).

47. Meldi, K. *et al*. Specific molecular signatures predict decitabine response in chronic myelomonocytic leukemia. *J. Clin. Invest.* **125**, 1857-1872 (2015).

48. Treppendahl, M.B., Kristensen, L.S. & Grønbæk, K. Predicting response to epigenetic therapy. *J. Clin. Invest.* **124**, 47-55 (2014).

49. Goodyear, O. *et al*. Induction of a CD8+ T-cell response to the MAGE cancer testis antigen by combined treatment with azacitidine and sodium valproate in patients with acute myeloid leukemia and myelodysplasia. *Blood* **116**, 1908-1918 (2010).

50. Yang, H. *et al.* Expression of PD-L1, PD-L2, PD-1 and CTLA4 in myelodysplastic syndromes is enhanced by treatment with hypomethylating agents. *Leukemia* **28**, 1280-1288 (2014).

51. Roulois, D. *et al*. DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts. *Cell* **162**, 961-973 (2015).

52. Chiappinelli, K. B. *et al*. Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. *Cell* **162**, 974-86 (2015).

53. Licht, J. D. DNA Methylation Inhibitors in Cancer Therapy: The Immunity Dimension. *Cell* **162**, 938-939 (2015).

54. Montenegro, M.F. *et al*. Targeting the epigenetic machinery of cancer cells. *Oncogene* **34**, 135-143 (2015).

55. Campbell, R.M. & Tummino, P.J. Cancer epigenetics drug discovery and development: the challenge of hitting the mark. *J. Clin. Invest.* **124**, 64-69 (2014).

56. Yoshida, K. *et al*. The landscape of somatic mutations in Down syndrome-related myeloid disorders. *Nat. Genet*. **45**, 1293-1299 (2013).

57. Koboldt, D.C. *et al*. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. **22**, 568-576 (2012).

58. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

59. Hua, X. *et al*. DrGaP: a powerful tool for identifying driver genes and pathways in cancer sequencing studies. *Am. J. Hum. Genet.* **93**, 439-451 (2013).

60. Larson, D.E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311-317 (2012).

61. Saunders, C.T. *et al*. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811-1817 (2012).

62. Fuentes Fajardo, K.V. *et al*. Detecting false-positive signals in exome sequencing. *Hum. Mutat.* **33**, 609-613 (2012).

63. Kim, D. *et al*. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).

64. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-359 (2012).

65. Anders, S., Pyl, P.T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**,166-169 (2015).

66. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol*. **15**, 550 (2014).

67. Park, Y., Figueroa, M.E., Rozek, L.S. & Sartor, M.A. MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics* **30**, 2414-2422 (2014).

68. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572 (2011).

END NOTES

Acknowledgments: This program was supported by grants from Ligue Nationale Contre le Cancer (équipe labellisée), Institut National du Cancer (INCa PLBIO, SIRIC SOCRATE), Agence Nationale de la Recherche (Molecular Medicine in Oncology ; Paris Alliance Cancer Research Institute : France Génomique National programs funded by « Investissements d'avenir »). JM was supported by the Fondation pour la Recherche Médicale (FDT20140931007).

Author contribution: JM, TQ, VG, MD, MJC, SK were involved in bioinformatics, ND in all the molecular analyses, ND, KY, NPM, SO in whole exome sequencing, ND, VM, FA, JFD in whole genome sequencing, ND, EC, DA, SS, DSB in RNA sequencing, MRS and LBA in mutational signature analysis, KM and MF in whole genome methylation, MM in patient sample manipulation and cell sorting, PF, TB, RI, SdB, BQ, EJ, SO, EP provided clinical samples, WV, OB, CP, EP discussed the results and provided advices, ES designed the study, analysed the results, and wrote the manuscript.

Conflicts of interest: No disclosure

Sequences were deposited at the European Genome-phenome Archive (EGA), hosted by the EBI, under accession number EGAS00001001264.

Legends

Figure 1. **Somatic variants in coding regions identified by whole exome sequencing (WES)**. WES was performed in 49 chronic myelomonocytic leukemia samples. (**a**) Number and type of somatic mutations identified in each patient designated as UPN, showing a majority of non synonymous variants. (**b**) Repartition of the 680 validated somatic variants identified in the 49 patients. (**c**) Repartition of base changes with transitions in black and transversions in grey. (**d**) Of the 36 recurrently mutated genes identified by WES, 26 are actively transcribed in $CD14^+$ cells and $CD34^+$ cells [according to Gene Expression Omnibus at http://www.ncbi.nlm.nih.gov/geo/]. These 26 recurrently mutated genes are classified according to their function, including epigenetic regulation, pre-messenger RNA splicing, and signal transduction. Colors indicate the type of mutation. Two colors separated by a slash indicate two distinct mutations in the same gene.

Figure 2. Somatic variants in coding and non-coding regions identified by whole genome sequencing (WGS). WGS was performed in 17 chronic myelomonocytic leukemia samples (including one analyzed by WES). (**a**) Number of somatic single nucleotide variants and short insertions/deletions in each patient. (**b**) Repartition of the 8077 somatic variants, expressed as numbers of variants per gigabase, identified across the genomic regions. (**c**) Repartition of base changes with transitions in black and transversions in grey. (**d**) Repartition of the 207 somatic variants identified in coding regions. (**e**) Mutational signatures extracted from whole genomic analyses. (**f)** Potential hotspots of mutations (2 variants less than 250 bp apart) including 9 in coding regions of driver genes (including *TET2*, *ASXL1*, *SRSF2*, *CBL* and *NRAS*), two in intronic regions of *PDS5A* and *NHLRC2*, one in 3'UTR of *ZFP36L2*, six in intergenic regions and 1 in the mitochondrial chromosome. Numbers between comas indicate the chromosome number.

Figure 3. TET3-R1548H mutation inhibits 5hmC modification. (**a)** Single cell analysis of TET3R1548H, TET2S1708fsX11 and TET2L1819X mutations in sorted $CD14^+$ cells from UPN22; (**b**) *TET2* and *TET3* gene expression measured by qRT-PCR in HEK293T cells transfected with the pcDNA3.1 empty vector or pcDNA3.1 encoding wildtype (TET3-WT) and R1548H TET3 (TET3-MUT). Reporter gene: *RPL32*. Results are related to pcDNA3.1

control. (**c**). Dot blot analysis of 5-hydroxymethylcytosine (5hmC) on genomic DNA (4-fold serial dilutions in ng) isolated from HEK293T cells transfected as in (**b**).

Figure 4. Serial whole exome sequencing (WES) analysis of somatic variants. WES of sorted peripheral blood monocyte DNA was performed 2- to 5-fold in 17 patients at a mean interval of 14+/-8 months (range 4-32). The clonal evolution of recurrently mutated genes is shown. UPN indicates the patient number. A selection of the variants detected by the first whole exome sequencing is shown (all the variants identified in each individual patient are depicted in supplementary figures 5 to 21). All the changes in variant allele frequency and new variants detected by repeating whole exome sequencing are shown. Black indicates the founding clone and subsequent subclones are shown in violet, red and orange successively. Patients were either untreated (**a**) or treated with either azacytidine (AZA) or decitabine (DAC) as indicated in red. Blue dash lines indicate WES. (**b**) patients with a stable disease on therapy; (**c**) responding patients.

Figure 5. Serial whole genome sequencing (WGS) in a 5-AZA exceptional responder. WGS was performed before 5-azatidine treatment (baseline), in complete response (remission) and at disease progression (relapse). (**a**) (**b**) Scatter plot of somatic variants identified at baseline, remission, and progression. Chromosomal location is color coded and the size of the object denotes its predicted impact on protein function. High impact variants are those that are predicted to have the highest likelihood of altering protein expression or function such as frameshifts or nonsense variants. Circles denote single nucleotide variants and triangles denote insertions or deletions. (**c**),(**d**) Scatter plot of all variants identified with Freebayes at baseline, remission, and progression. Chromosomal location is color-coded. (**e**) Copy number changes as identified from whole genome sequencing data using Sequenza.

Figure 6. **Evolution of gene expression pattern upon hypomethylating agent therapy.** Gene expression was analyzed at two time points in sorted peripheral blood monocytes from 9 chronic myelomonocytic leukemia patients, including 3 untreated and 6 treated with either azacytidine or decitabine. These cases were randomly selected in each group. Three treated patients remained stable on therapy (non-responders) whereas the

three others were responders. In treated patients, the first sample was collected before treatment, the second one after at least 5 drug cycles and just before the next cycle. Volcano plots of genes differentially expressed between these two time points are shown in non-responders (**a**) and in responders (**b**). The name of the most differentially deregulated genes is indicated. No significant change in gene expression was detected in untreated patients analyzed twice at an at least 5 month-interval (see also **table 1**). Each dot (N=24,563) represents a gene; green dots, padj<=0.05, orange dots, abs(log2FoldChange)>=1 and red dots, padj<=0.05 and abs(log2FoldChange)>=1. (**c**) qRT-PCR validation of the differential expression of 8 genes in 6 responders [3 studied by RNA sequencing in (b) and 3 additional cases] and 10 non-responders [3 studied by RNA sequencing in (a) and 7 additional cases]. Normalizer gene, *RPL32*. Similar results were obtained with 2 other normalizer genes, *GUS* and *HPRT* (**supplementary figure 24**). (**d**) Significant changes in pathways detected by analyzing RNA sequencing data with Ingenuity (www.ingenuity.com/products/ipa).

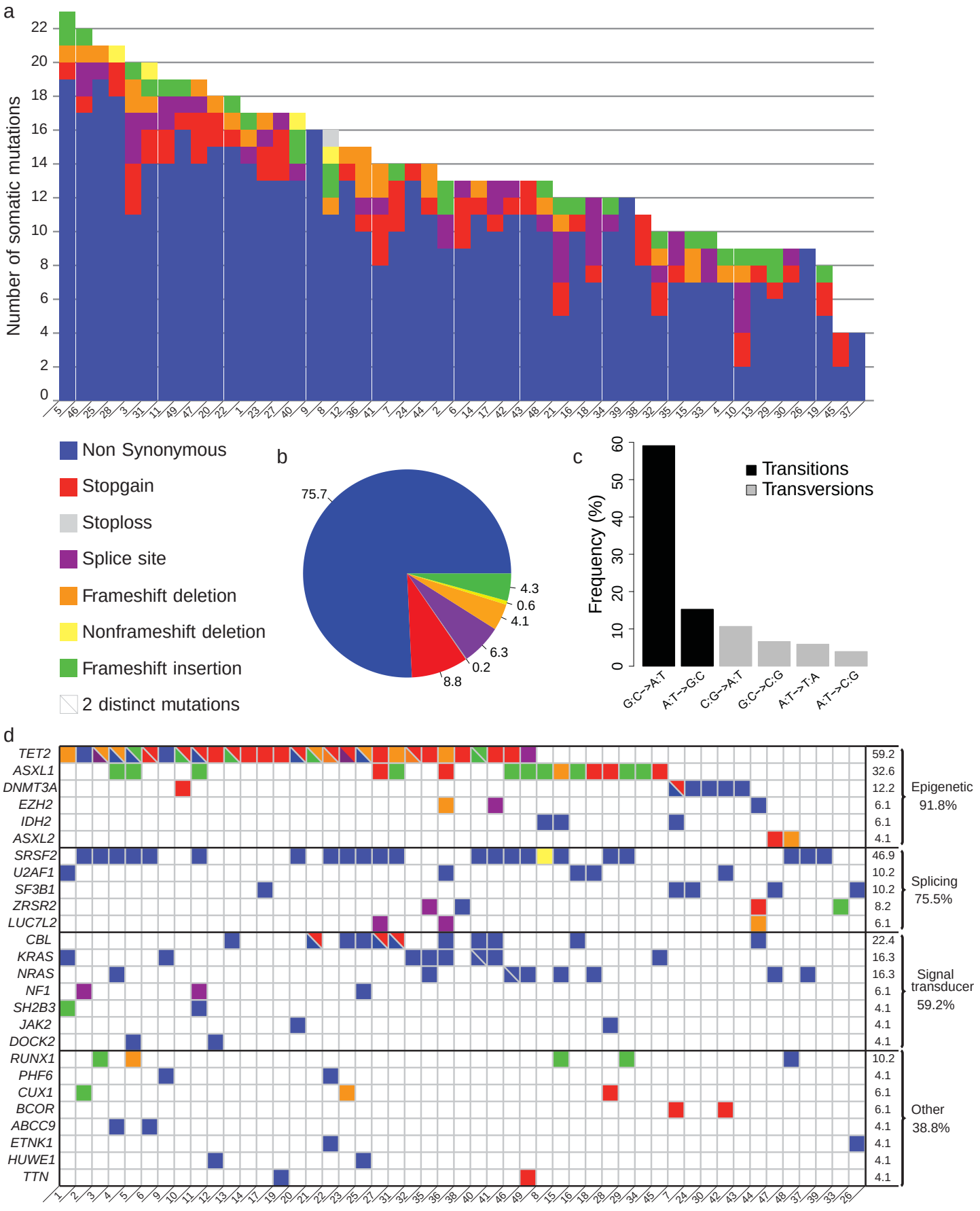
Figure 7. Evolution of DNA methylation pattern upon hypomethylating drug therapy. Methylation was analyzed at two time points in sorted monocytes from 9 chronic myelomonocytic leukemia patients, including 3 untreated and 6 treated with either azacytidine or decitabine. Three treated patients remained stable on therapy (non-responders) whereas the three others were responders. In treated patients, the first sample was collected before treatment, the second one after at least 5 drug cycles and just before the next cycle. **a,b.** Chromosome ideograms representing differentially methylated regions (DMRs) in non-responders (**a**) and in responders (**b**) are shown. Reduction in DNA methylation is in green, whereas increased methylation is in pink. **c.d.** Barplots showing the percentage of genomic regions with significant changes in DNA methylation in non-responders (**c**) and in responders (**d**) are also shown. No change was identified in the 3 untreated patients (**Table 1**). **e**. Violin plots showing the evolution of global methylation change in each patient (untreated patients in grey, treated with a stable disease (non-responders) in blue, treated responders in red with the lighter color indicating the earliest analysis.

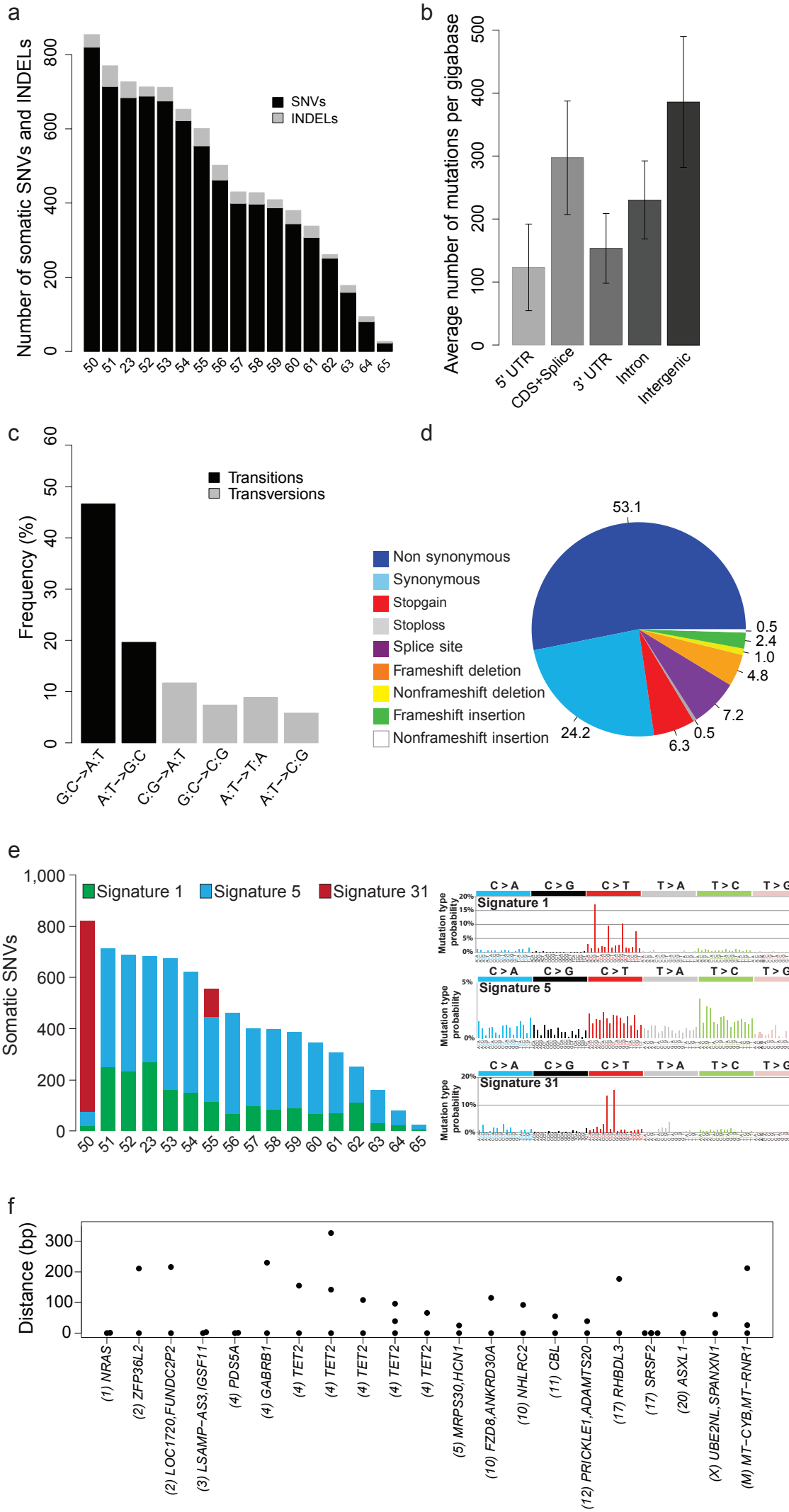
Figure 8. Relationship between changes in DNA methylation and in gene expression. Venn diagrams of interactions between differentially methylated regions (DMR, red circles) and differentially expressed genes (up-regulated in blue; down regulated in green) as defined in figure 5 (padj<=0.05 and abs(log2FoldChange)>=1).

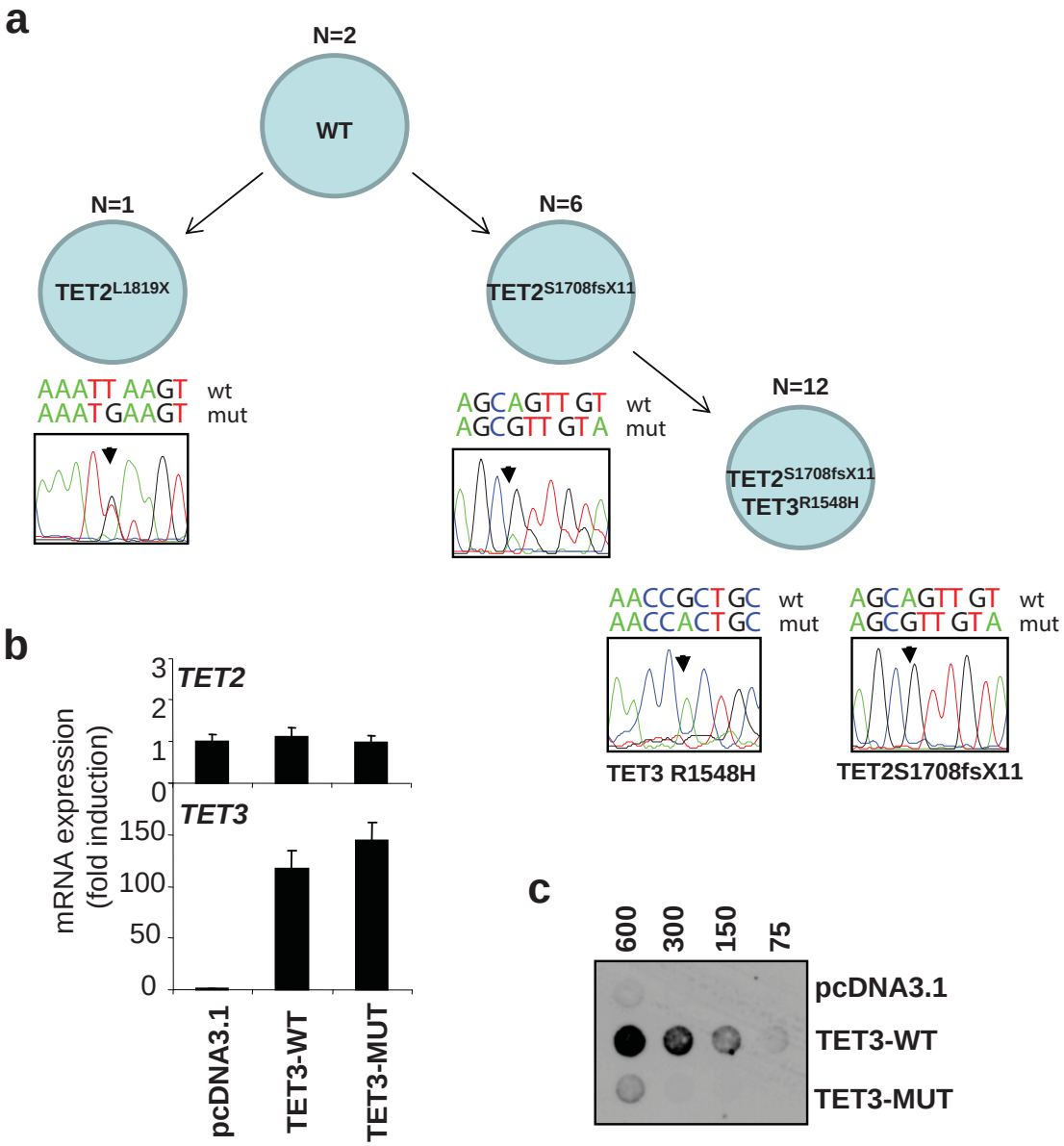
Table 1. **Changes induced by hypomethylating agents in gene expression and DNA methylation.**

		Untreated	Treated Non-responders	Treated Responders
Number of patients		3	3	3
Time between analyses (months)		17 ± 10	9 ± 3	20 ± 4
Changes in gene expression	Up	0	12	343
	Down	0	51	170
	Total	**0**	**63**	**513**
Differentially methylated regions	Up	0	28	19
	Down	1	75	35,895
	Total	**0**	**103**	**35,914**

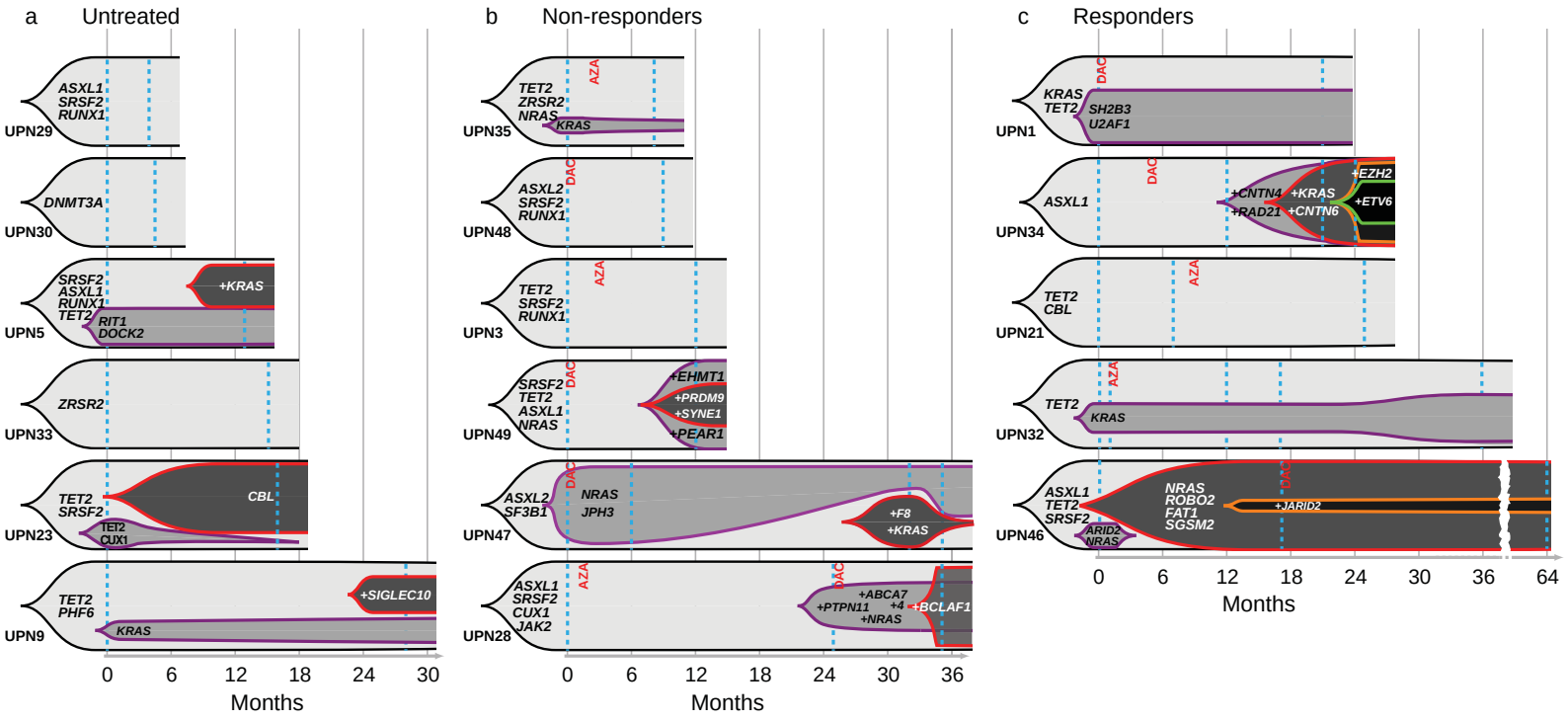
Genomic analyses were performed at two time points in sorted peripheral blood monocytes of 9 chronic myelomonocytic leukemia patients, including 3 left untreated and 6 patients treated with either azacytidine or decitabine. Among treated patients, 3 had a stable disease under therapy (non-responders) and 3 demonstrated clinical response (**Fig. 4** and **5**). The first sample was collected before treatment, the second after at least 5 cycles of either azacytidine or decitabine, just before the next cycle. We measured the number of differentially expressed genes having abs(log2FoldChange)≥1 between T1 and T2, and the number of differentially methylated regions having≥25% difference between T1 and T2.

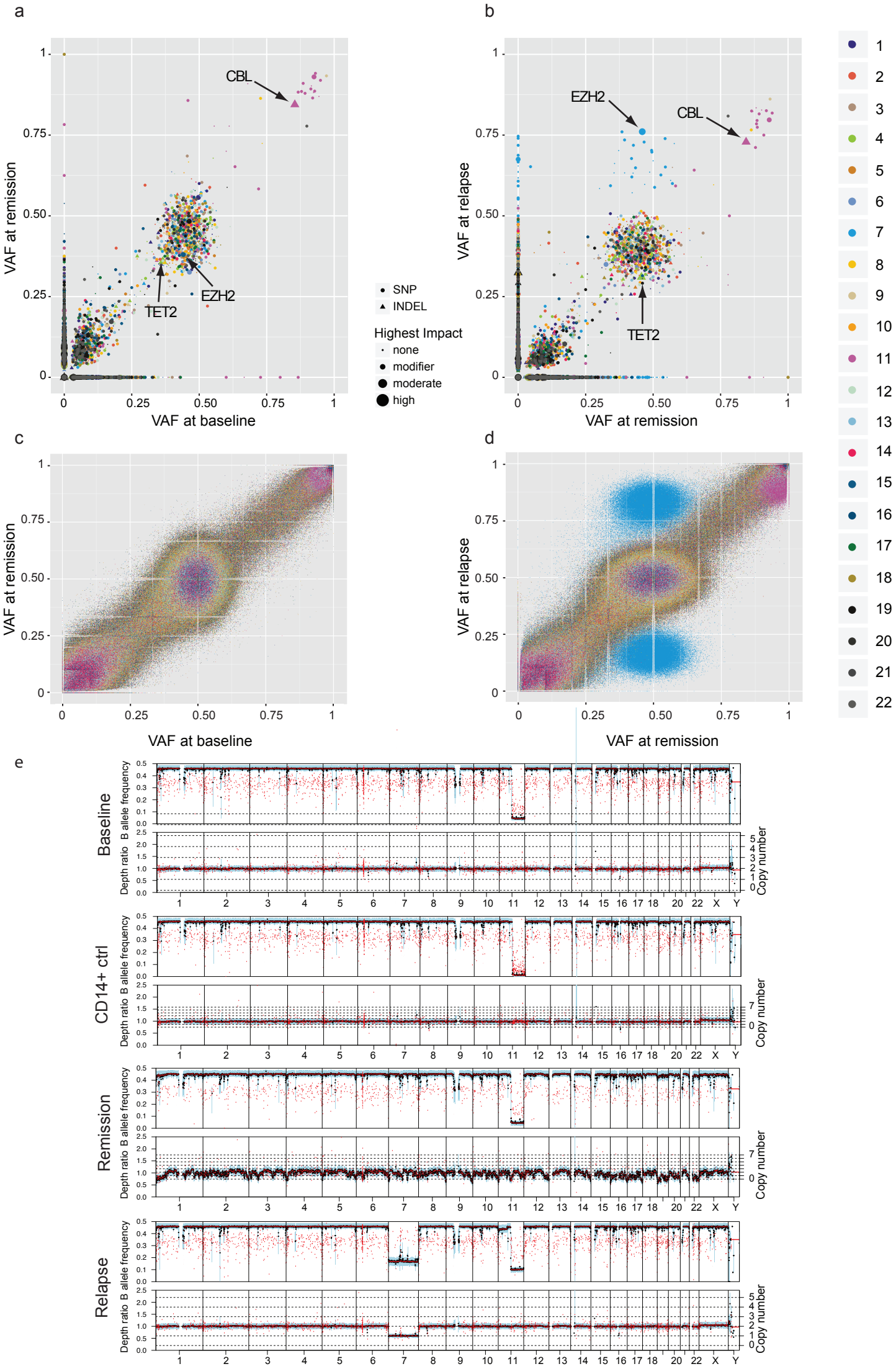

a
Number of somatic mutations
0
2
4
6
8
10
12
14
16
18
20
22
5 46 25 28 3 31 11 49 47 20 22 1 23 27 40 9 8 12 36 41 7 24 44 2 6 14 17 42 43 48 21 16 18 34 39 38 32 35 15 33 4 10 13 29 30 26 19 45 37
Non Synonymous
Stopgain
Stoploss
Splice site
Frameshift deletion
Nonframeshift deletion
Frameshift insertion
2 distinct mutations
b
75.7
4.3
0.6
4.1
6.3
0.2
8.8
c
Frequency (%)
0
10
20
30
40
50
60
Transitions
Transversions
G:C->A:T
A:T->G:C
C:G->A:T
G:C->C:G
A:T->T:A
A:T->C:G
d
TET2 59.2
ASXL1 32.6
DNMT3A 12.2
EZH2 6.1
IDH2 6.1
ASXL2 4.1
Epigenetic 91.8%
SRSF2 46.9
U2AF1 10.2
SF3B1 10.2
ZRSR2 8.2
LUC7L2 6.1
Splicing 75.5%
CBL 22.4
KRAS 16.3
NRAS 16.3
NF1 6.1
SH2B3 4.1
JAK2 4.1
DOCK2 4.1
Signal transducer 59.2%
RUNX1 10.2
PHF6 4.1
CUX1 6.1
BCOR 6.1
ABCC9 4.1
ETNK1 4.1
HUWE1 4.1
TTN 4.1
Other 38.8%
1 2 3 4 5 6 9 10 11 12 13 14 17 19 20 21 22 23 25 27 31 32 35 36 38 40 41 46 49 8 15 16 18 28 29 34 45 7 24 30 42 43 44 47 48 37 39 33 26


a
Number of somatic SNVs and INDELs
SNVs
INDELs
50 51 23 52 53 54 55 56 57 58 59 60 61 62 63 64 65
b
Average number of mutations per gigabase
5' UTR
CDS+Splice
3' UTR
Intron
Intergenic
c
Frequency (%)
Transitions
Transversions
G:C->A:T
A:T->G:C
C:G->A:T
G:C->C:G
A:T->T:A
A:T->C:G
d
Non synonymous
Synonymous
Stopgain
Stoploss
Splice site
Frameshift deletion
Nonframeshift deletion
Frameshift insertion
Nonframeshift insertion
53.1
24.2
6.3
0.5
7.2
4.8
1.0
2.4
0.5
e
Somatic SNVs
Signature 1
Signature 5
Signature 31
C > A
C > G
C > T
T > A
T > C
T > G
Mutation type probability
f
Distance (bp)
(1) NRAS
(2) ZFP36L2
(2) LOC1720,FUNDC2P2
(3) LSAMP-AS3,IGSF11
(4) PDS5A
(4) GABRB1
(4) TET2
(4) TET2
(4) TET2
(4) TET2
(4) TET2
(5) MRPS30,HCN1
(10) FZD8,ANKRD30A
(10) NHLRC2
(11) CBL
(12) PRICKLE1,ADAMTS20
(17) RHBDL3
(17) SRSF2
(20) ASXL1
(X) UBE2NL,SPANXN1
(M) MT-CYB,MT-RNR1


a
N=2
WT
N=1
TET2L1819X
N=6
TET2S1708fsX11
AAATT AAGT wt
AAAT GAAGT mut
AGCAGTT GT wt
AGCGTT GTA mut
N=12
TET2S1708fsX11
TET3R1548H
AACCGCTGC wt
AACCACTGC mut
TET3 R1548H
AGCAGTT GT wt
AGCGTT GTA mut
TET2S1708fsX11
b
mRNA expression (fold induction)
TET2
TET3
pcDNA3.1
TET3-WT
TET3-MUT
c
600
300
150
75
pcDNA3.1
TET3-WT
TET3-MUT

Merlevede et al, Figure 4





Merlevede et al, Figure 5

a

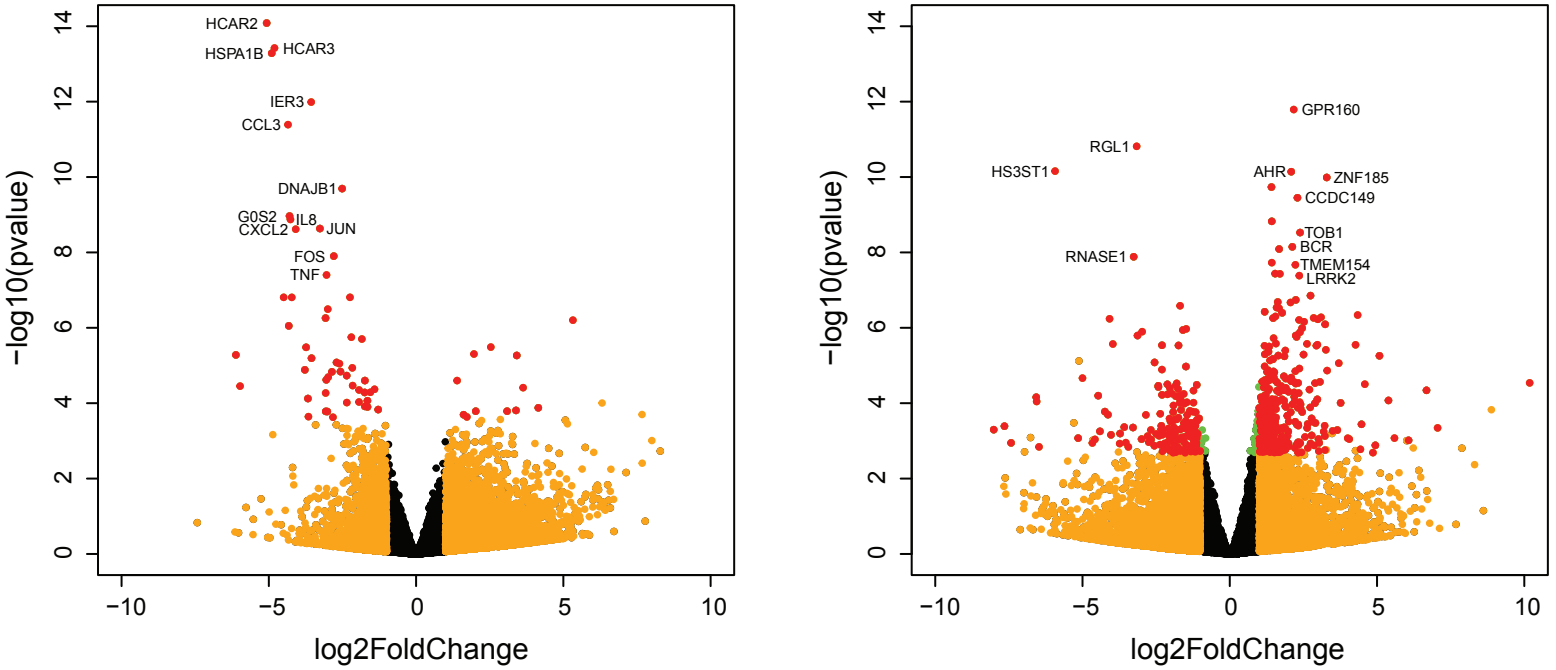
Non-responders

(UPN47, UPN48, UPN49)

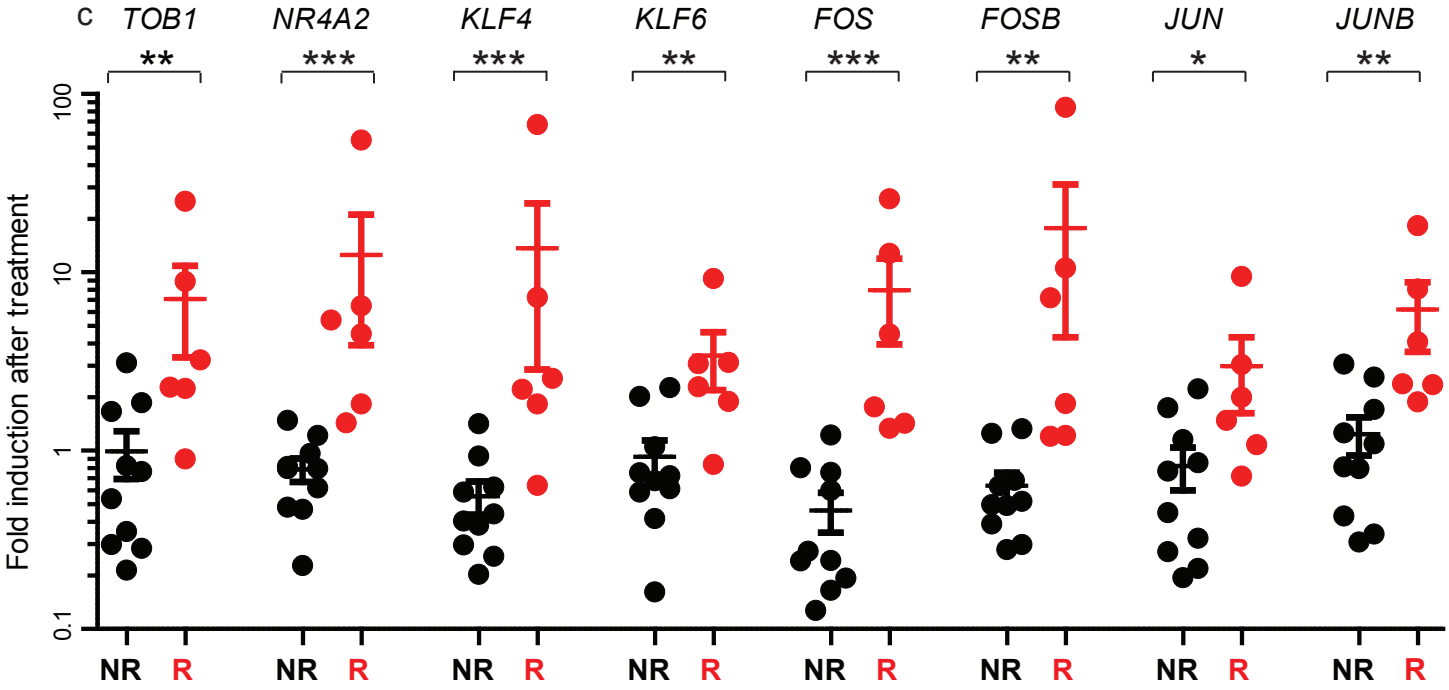
b

Responders

(UPN21, UPN32, UPN46)



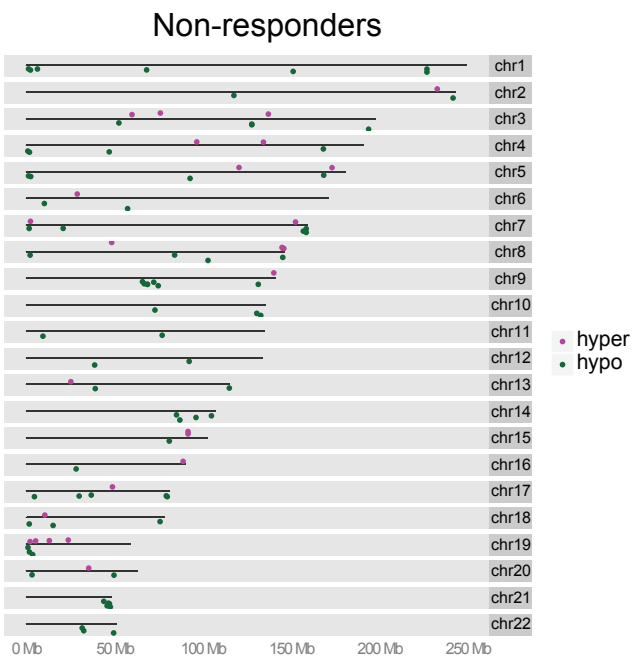
c



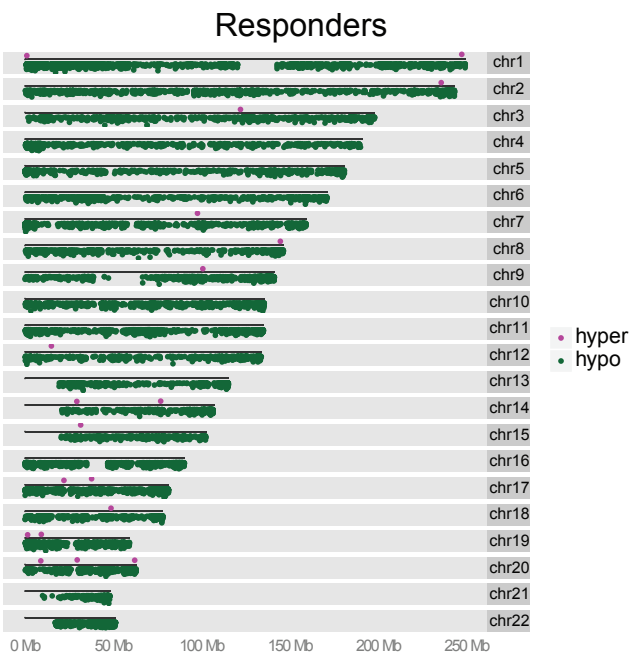
d

Upstream regulators	p-value	Predicted activation
Lipopolysaccharide	$3.84\ 10^{-13}$	Activated
IFN-gamma	$1.49\ 10^{-11}$	Activated
CD40 ligand	$2.52\ 10^{-10}$	Activated
IL1 beta	$1.03\ 10^{-9}$	Activated

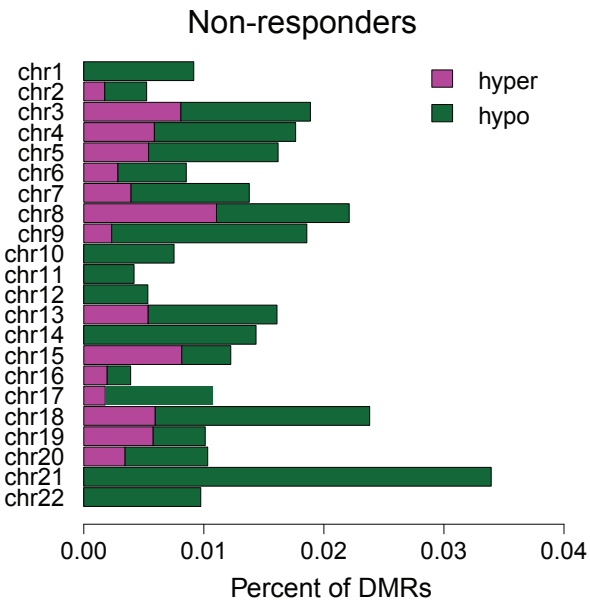
a


Non-responders
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
0 Mb
50 Mb
100 Mb
150 Mb
200 Mb
250 Mb
hyper
hypo

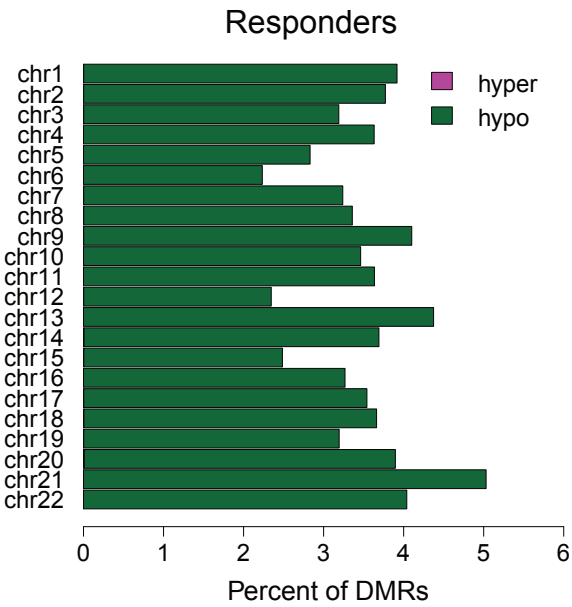
b


Responders
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
0 Mb
50 Mb
100 Mb
150 Mb
200 Mb
250 Mb
hyper
hypo

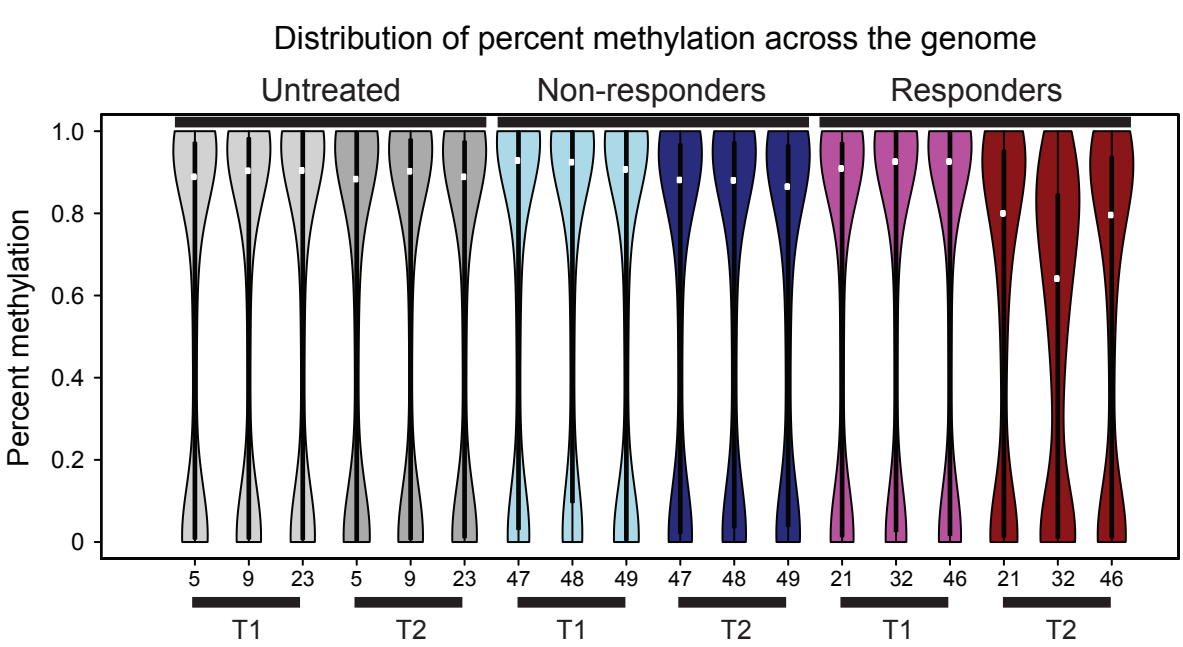
c


Non-responders
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
hyper
hypo
0.00
0.01
0.02
0.03
0.04
Percent of DMRs

d


Responders
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
hyper
hypo
0
1
2
3
4
5
6
Percent of DMRs

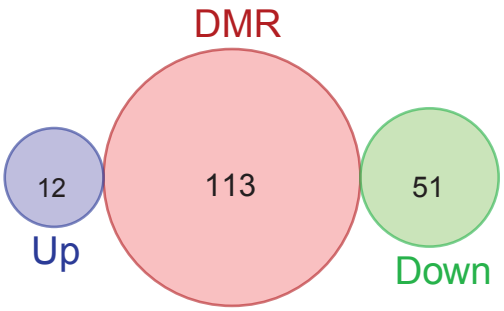
e


Distribution of percent methylation across the genome
Untreated
Non-responders
Responders
Percent methylation
1.0
0.8
0.6
0.4
0.2
0
5
9
23
5
9
23
47
48
49
47
48
49
21
32
46
21
32
46
T1
T2
T1
T2
T1
T2

Merlevede et al, Figure 8

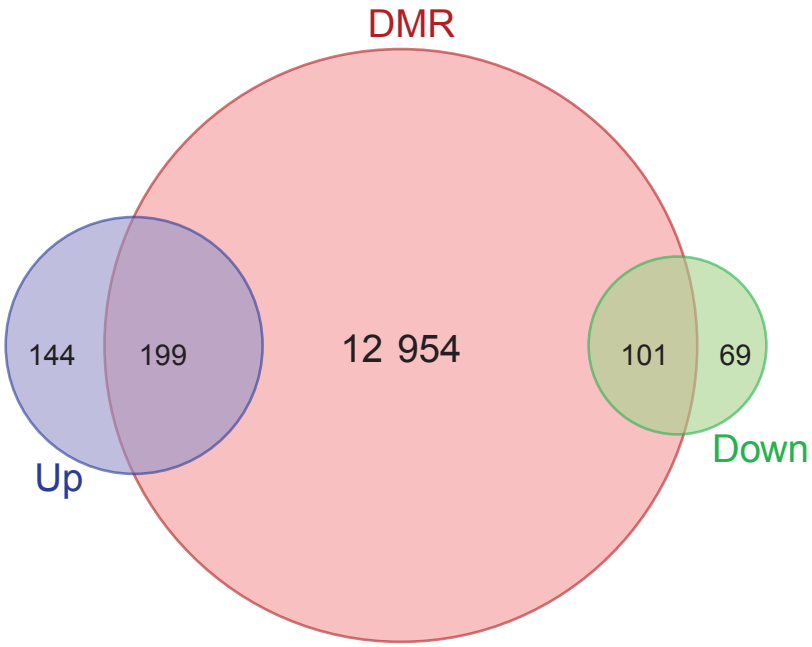
Non-responders

(UPN47, UPN48, UPN49)



Responders

(UPN21, UPN32, UPN46)



Supplementary Appendix

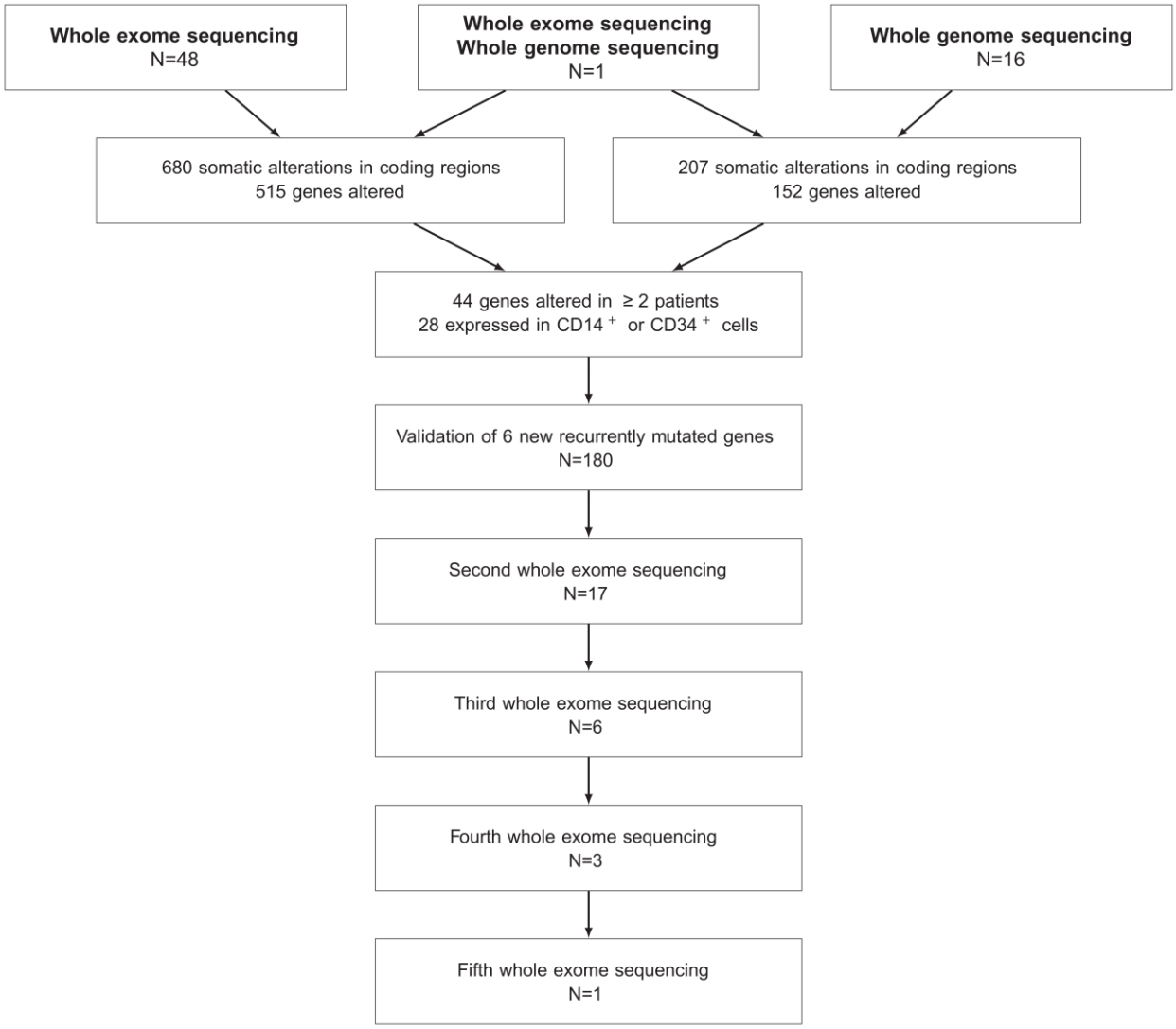
Mutation allele burden remains unchanged in chronic myelomonocytic leukemia responding to hypomethylating agents.

Jane Merlevede, Nathalie Droin, Tingting Qin, Kristen Meldi, Kenichi Yoshida, Margot Morabito, Emilie Chautard, Didier Auboeuf, Pierre Fenaux, Thorsten Braun, Raphael Itzykson, Stéphane de Botton, Bruno Quesnel, Thérèse Commes, Eric Jourdan, William Vainchenker, Olivier Bernard, Noemie Pata-Merci, Stéphanie Solier, Velimir Gayevskiy, Marcel E Dinger, Mark J Cowley, Dorothée Selimoglu-Buet, Vincent Meyer, François Artiguenave, Jean-François Deleuze, Claude Preudhomme, Michael R Stratton, Ludmil B Alexandrov, Eric Padron, Seishi Ogawa, Serge Koscielny, Maria Figueroa, Eric Solary.

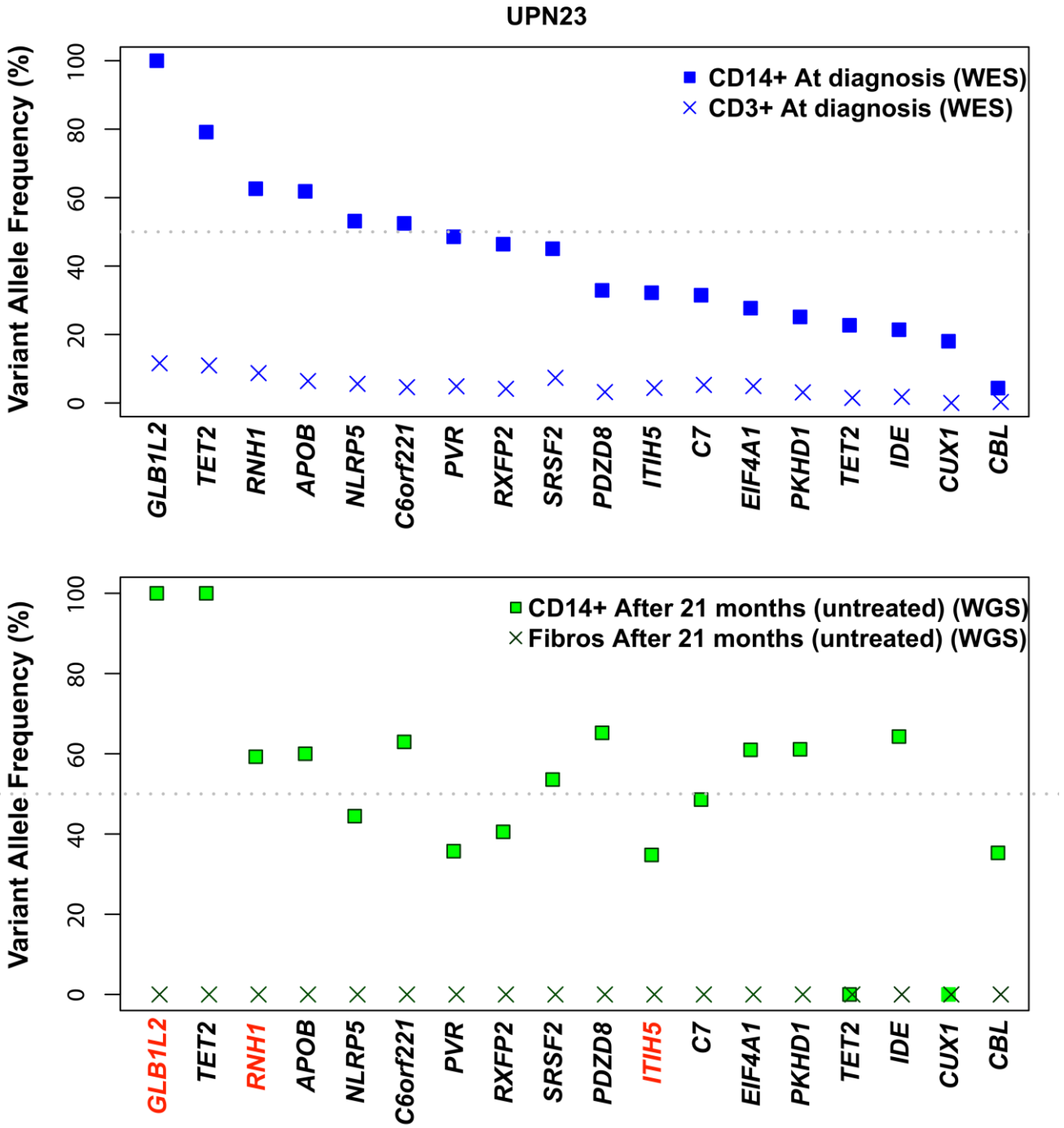
Table of Contents

I - Supplementary Figures

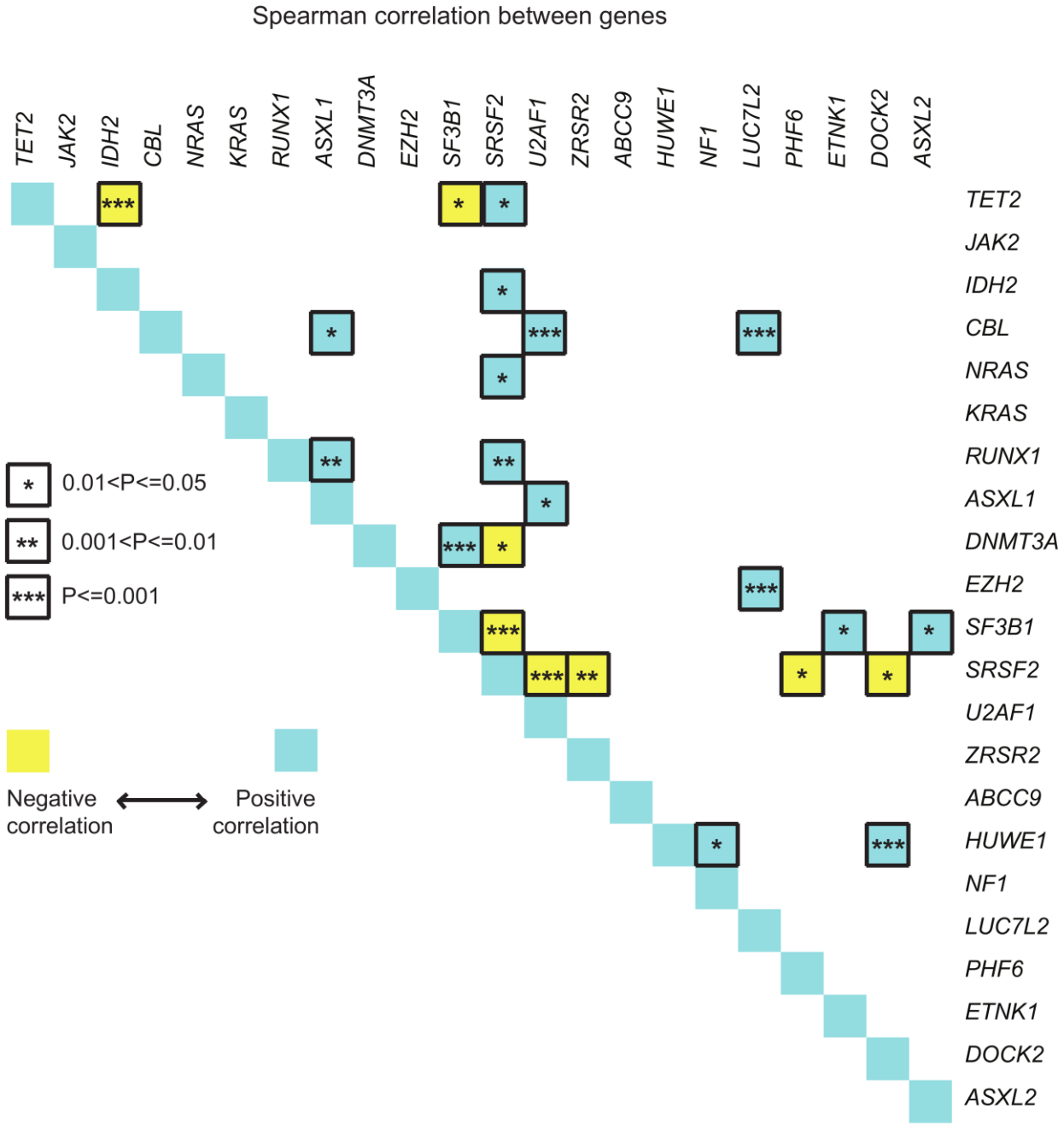
Supplementary Figure 1 - Flowchart of the realized experiments



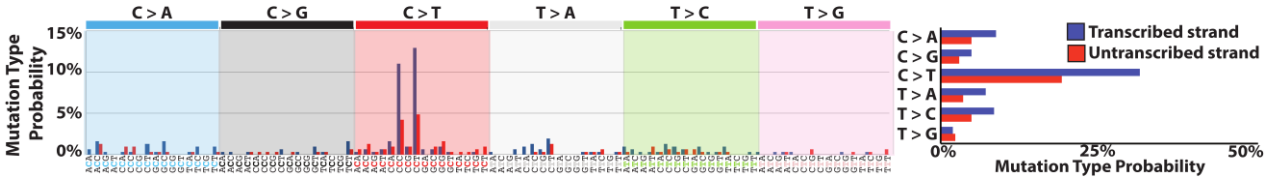
Supplementary Figure 2: Comparison of skin fibroblasts and CD3+ cells as germline controls. In UPN 23, we used CD3+ and skin fibroblasts as controls of whole exome and whole genome sequencing of $CD14^{+}$ sorted monocyte DNA, respectively. In both analyses, the same abnormalities were detected in the coding regions of monocyte DNA. More variation in variant allele frequencies was observed in whole genome sequences, due to lower coverage.



Supplementary Figure 3: Positive and negative correlations between recurrent gene mutations



Supplementary Figure 4: Characteristics of mutational signature 31. The new signature 31 is characterized by C:G>T:A mutations at CpCpC and CpCpT (mutated based underlined) and exhibits a strong transcriptional strand bias (especially in regards to C:G>T:A mutations, with mutations occurring predominately on guanine) as illustrated below.



Supplementary figures 5 to 21: Allele frequency of the variants detected by serial whole exome sequencing of sorted CD14+ monocyte DNA in 17 patients. We did not detect any change in the number of variants by serial analyses shown on supplementary figures 5 to 8 and 10 to 15, including 4 untreated patients (29,30,33,23), 3 non responding patients with a stable disease upon treatment with a demethylating agent (35,48,3) and 3 patients who responded to treatment with a demethylating agent (32,1,21). We observed changes in the number of variants by serial analyses shown on supplementary figures 9, 16 to 21, including 3 untreated patients [5, 9 and 46 (46 being treated afterwards)], 3 so-called "non-responding" patients with a stable disease upon treatment with a demethylating agent (47,49,28) and one patient who responded to the demethylating drug (34).

Figure S5: Serial whole exome sequencing in patient 29

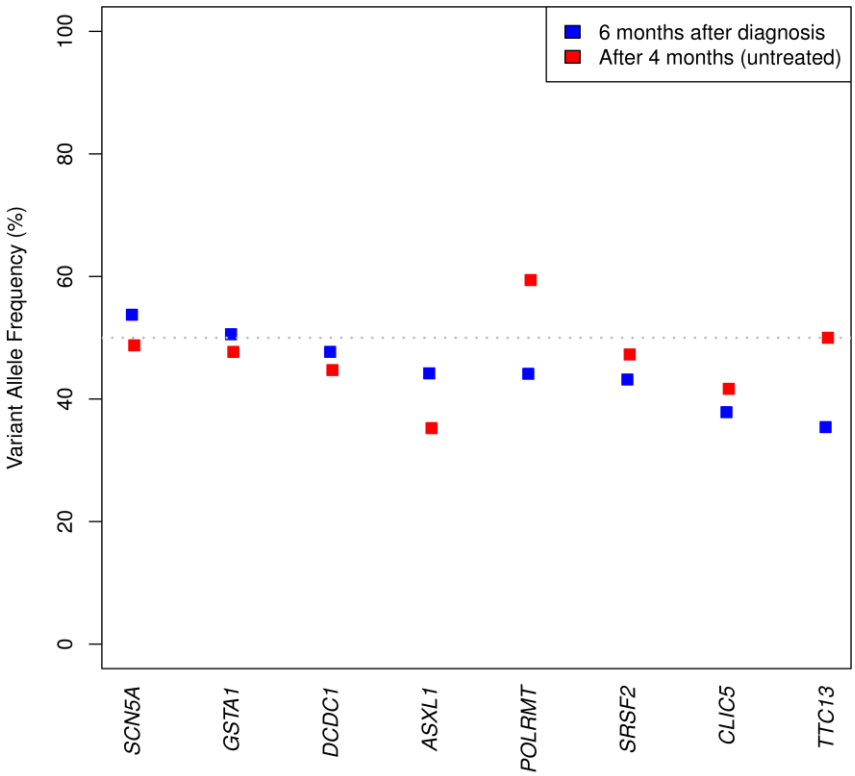


Figure S6: Serial whole exome sequencing in patient 30

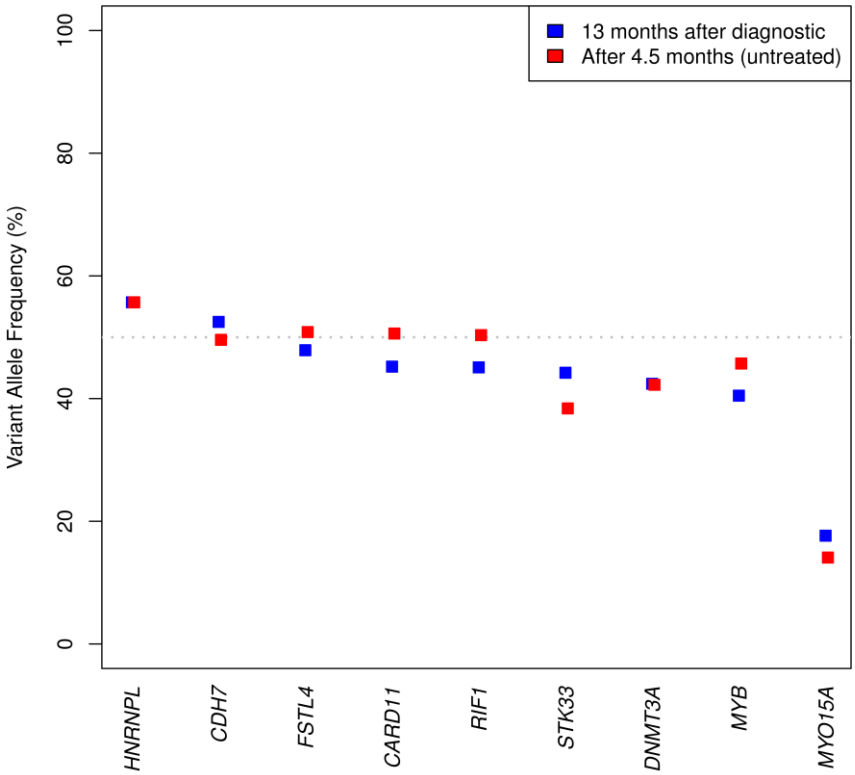


Figure S7: Serial whole exome sequencing in patient 33

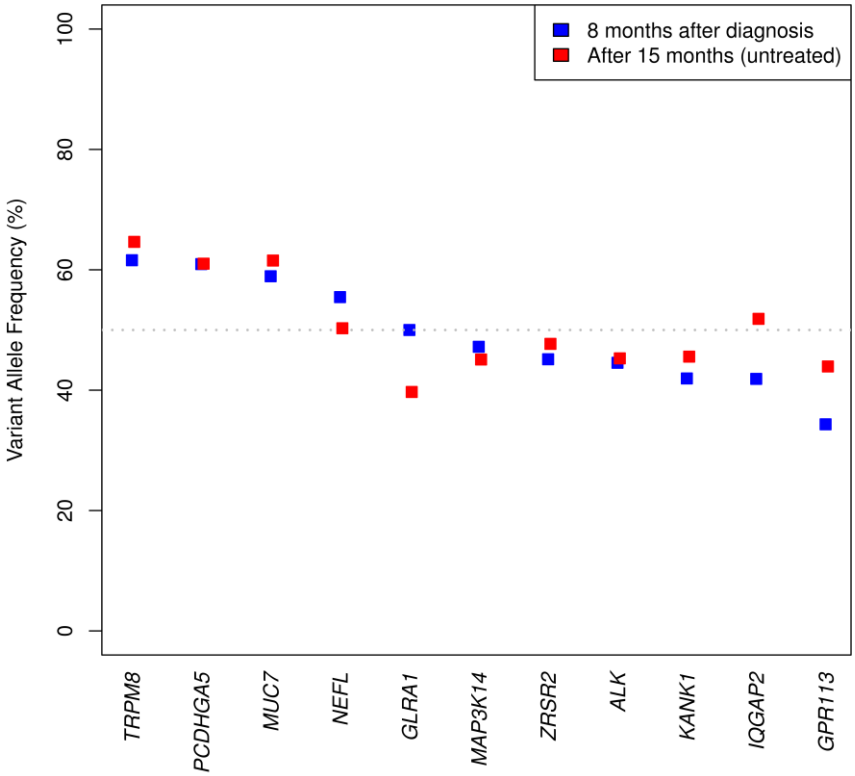


Figure S8: Serial whole exome sequencing in patient 23

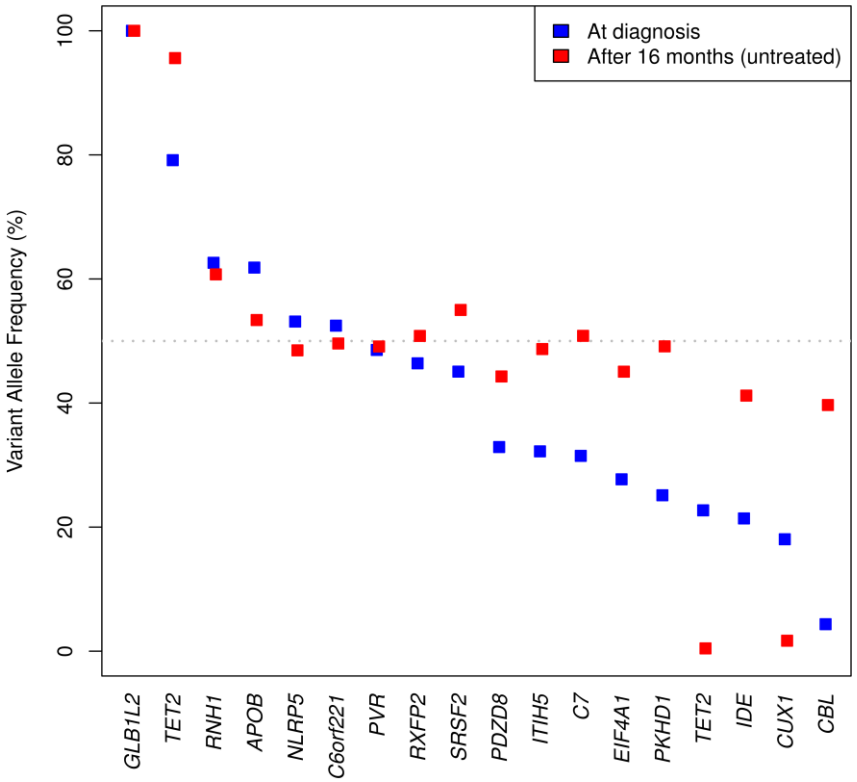


Figure S9: Serial whole exome sequencing in patient 47

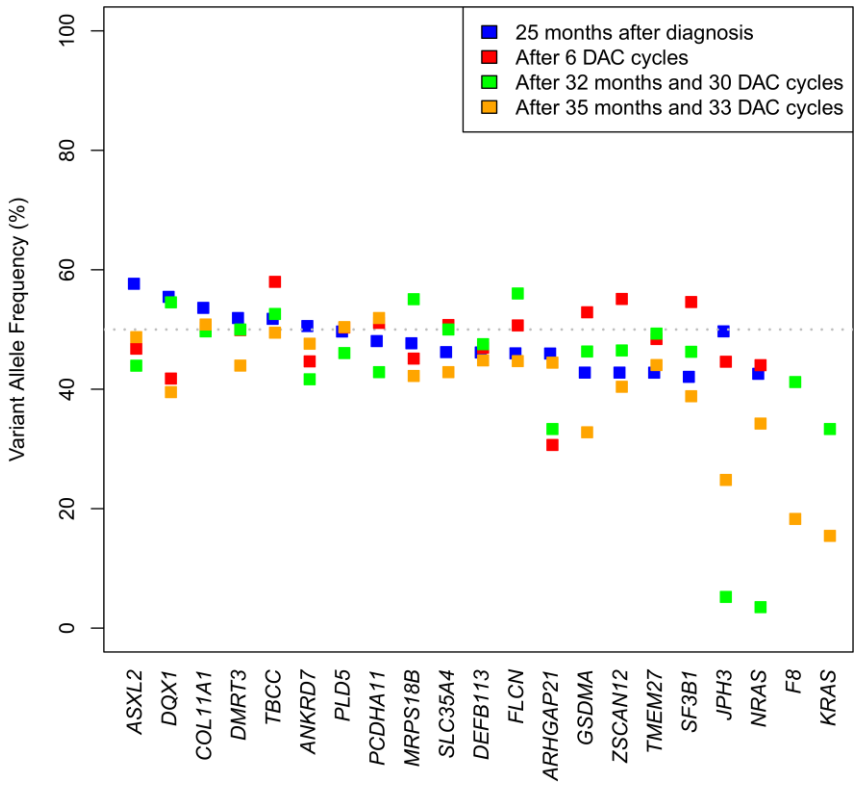


Figure S10: Serial whole exome sequencing in patient 35

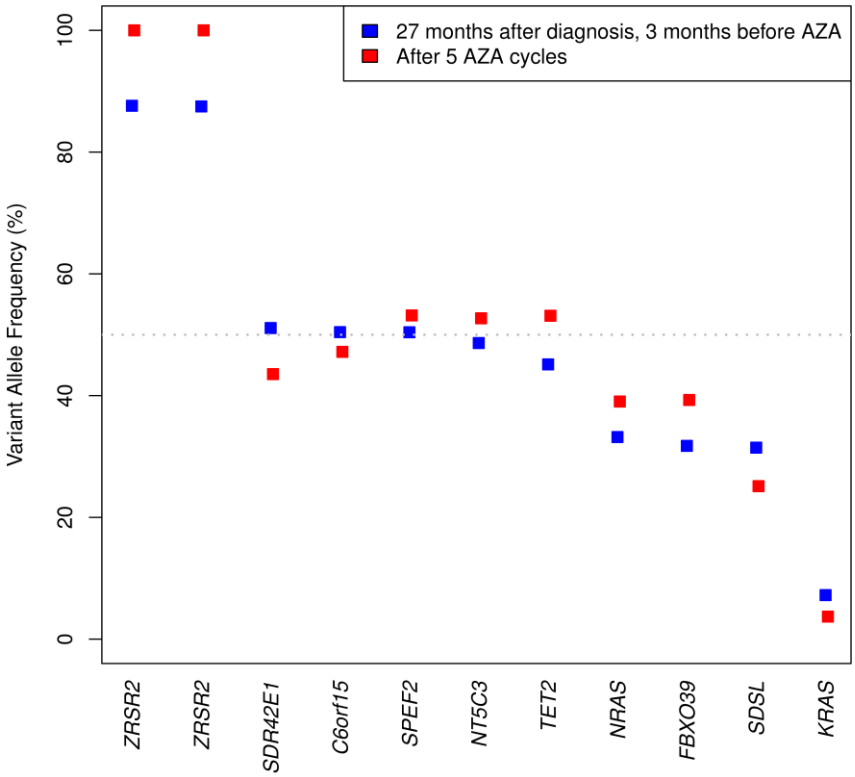


Figure S11: Serial whole exome sequencing in patient 48

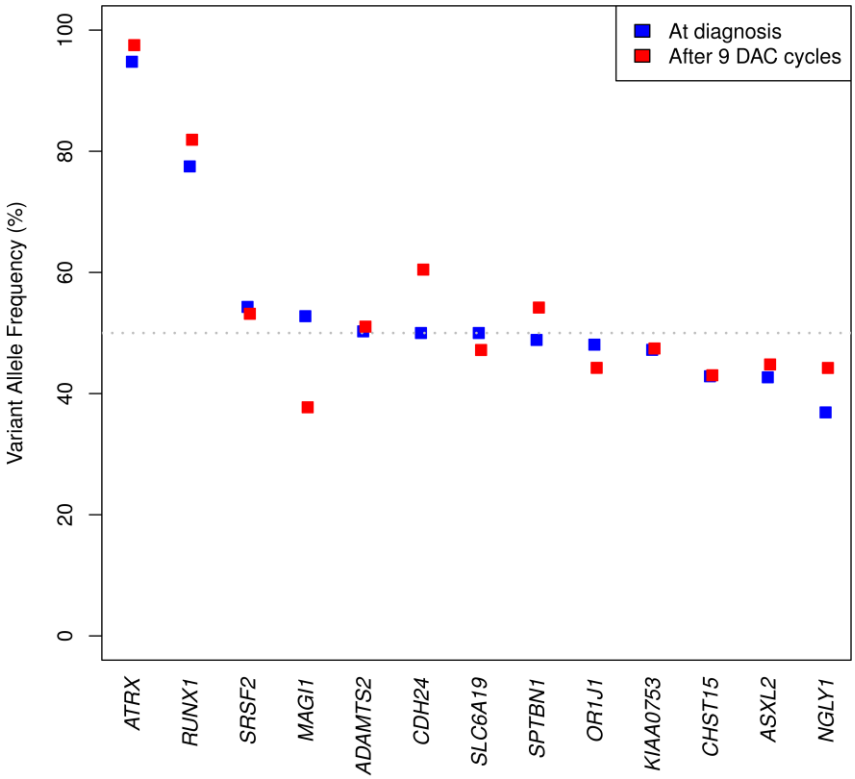


Figure S12: Serial whole exome sequencing in patient 3

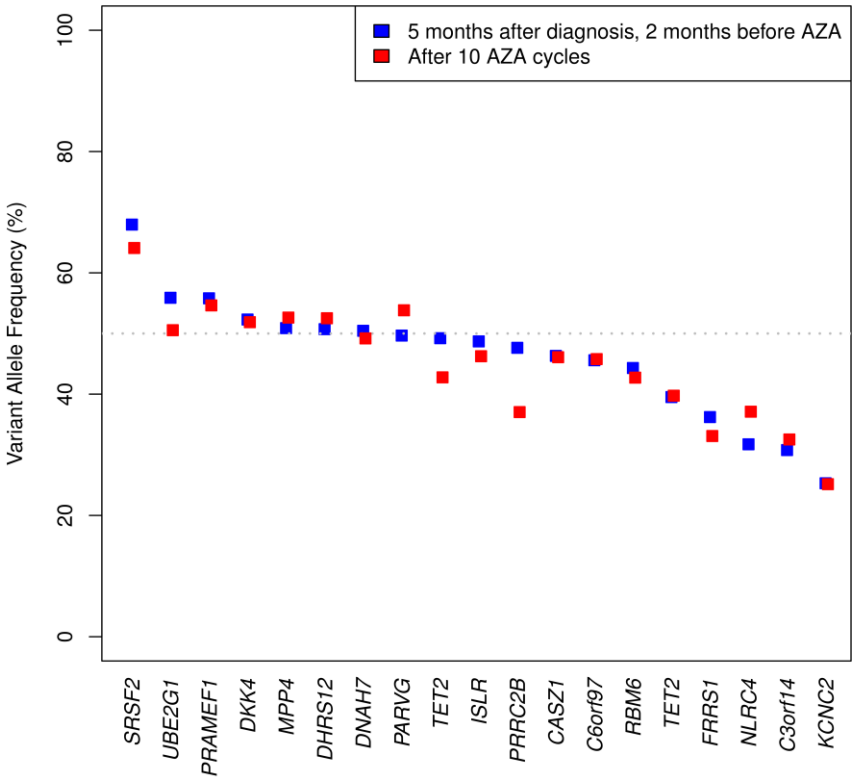


Figure S13: Serial whole exome sequencing in patient 32

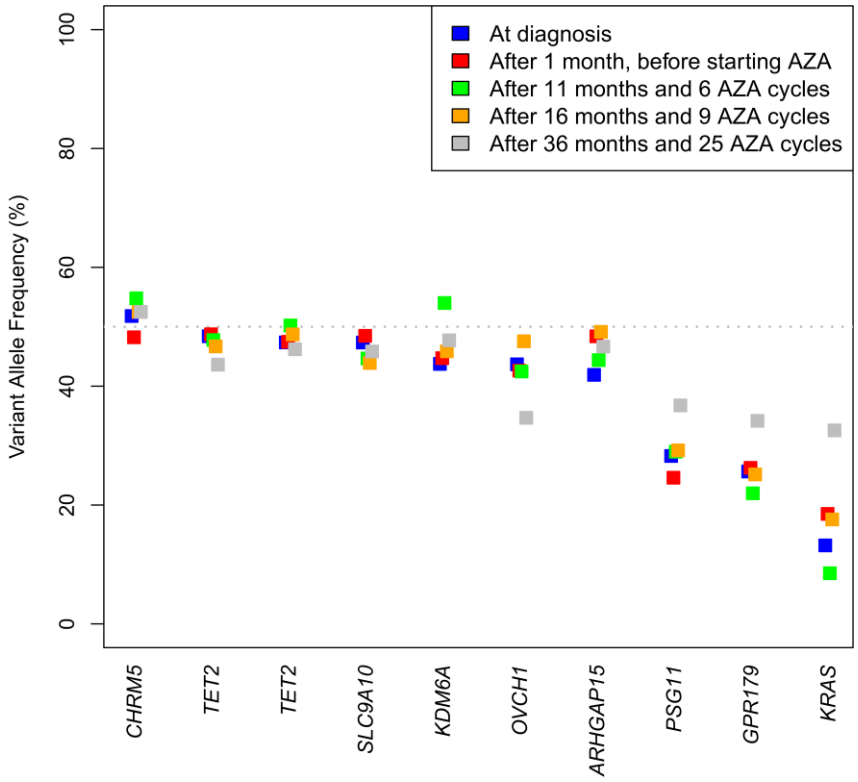


Figure S14: Serial whole exome sequencing in patient 1

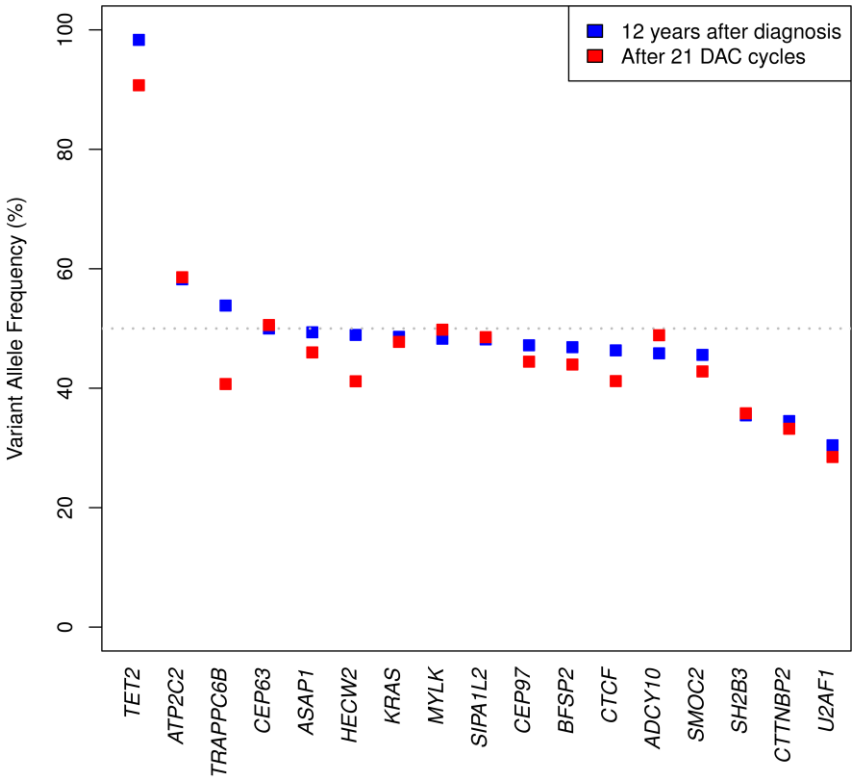


Figure S15: Serial whole exome sequencing in patient 21

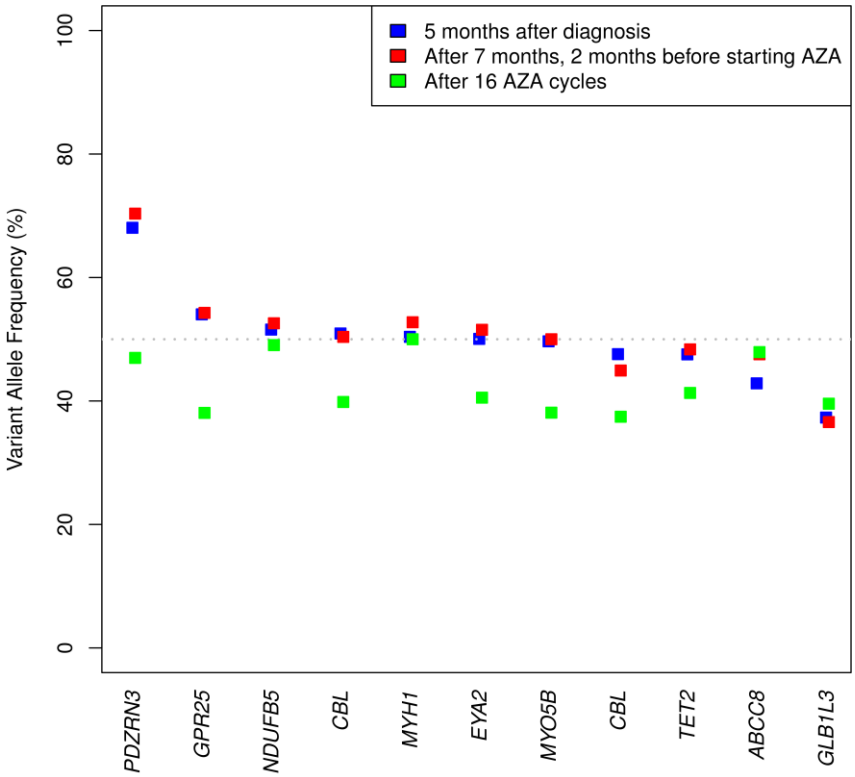


Figure S16: Serial whole exome sequencing in patient 5

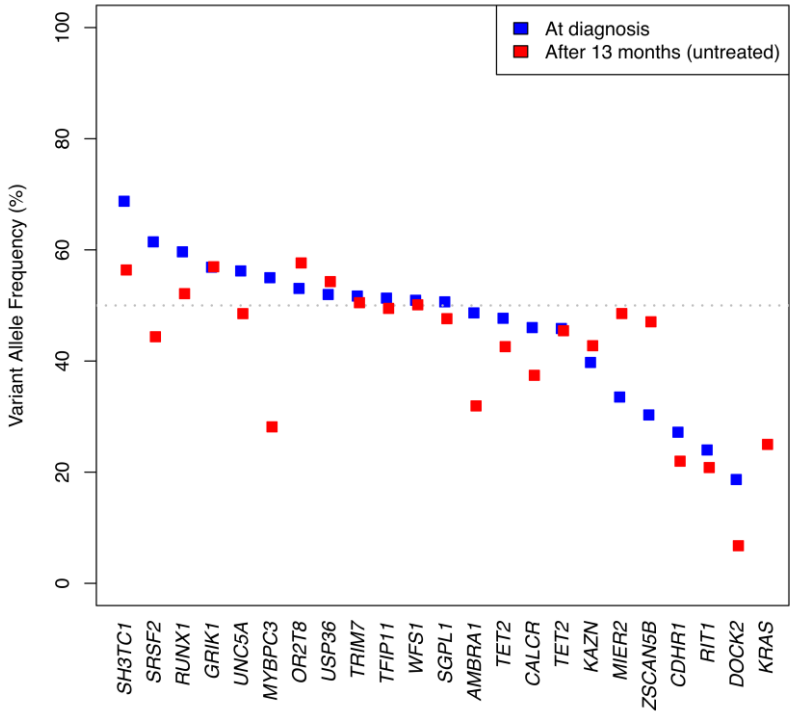


Figure S17: Serial whole exome sequencing in patient 9

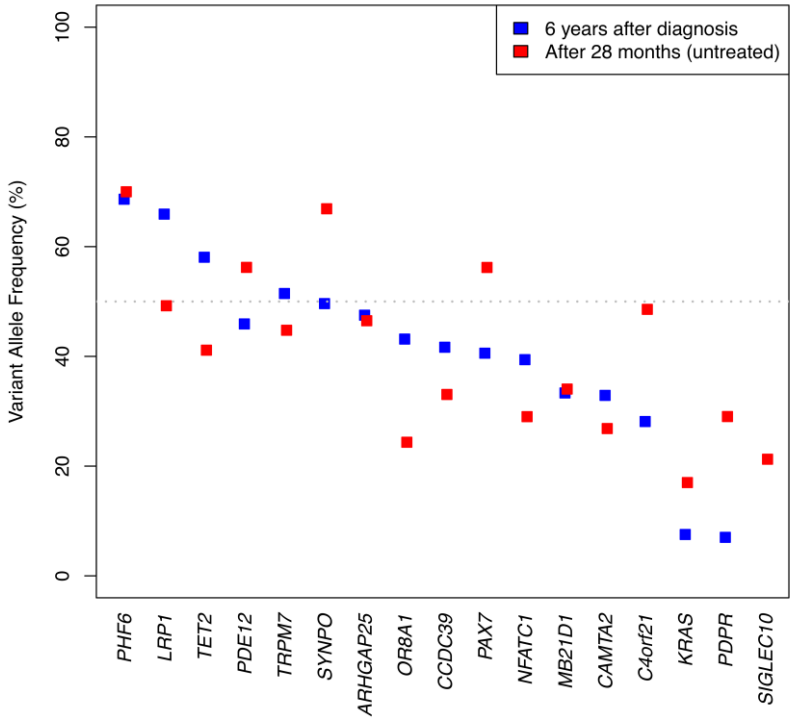


Figure S18: Serial whole exome sequencing in patient 46

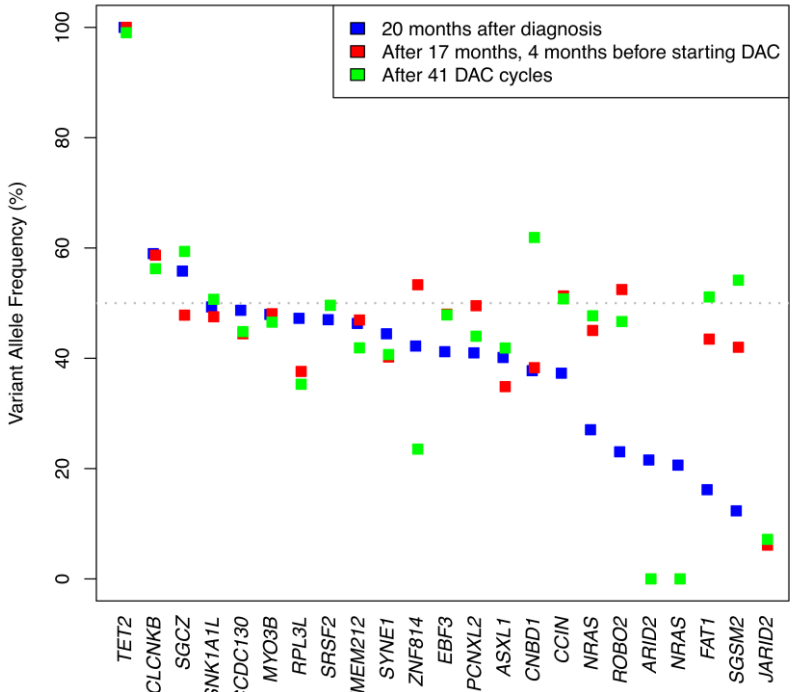


Figure S19: Serial whole exome sequencing in patient 49

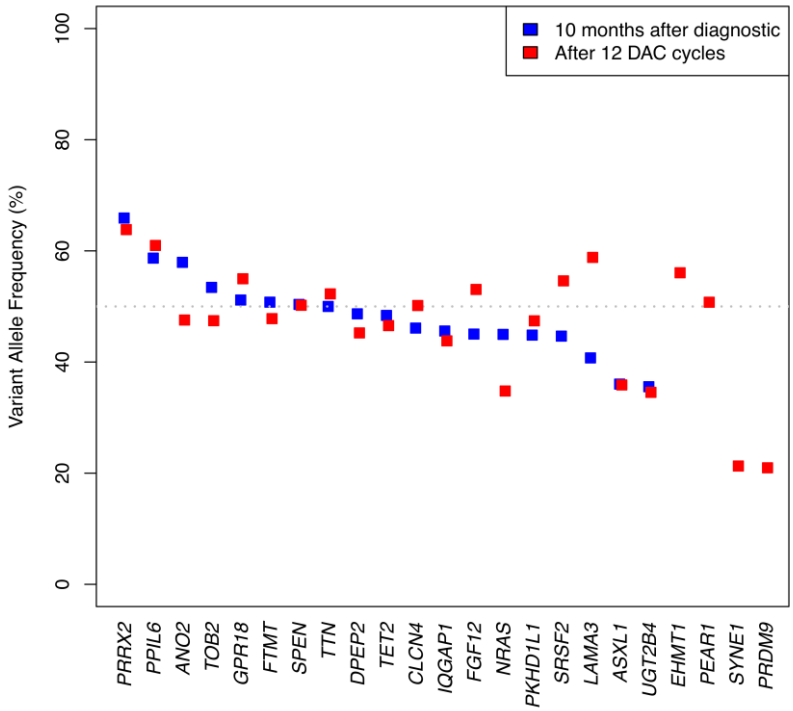


Figure S20: Serial whole exome sequencing in patient 28

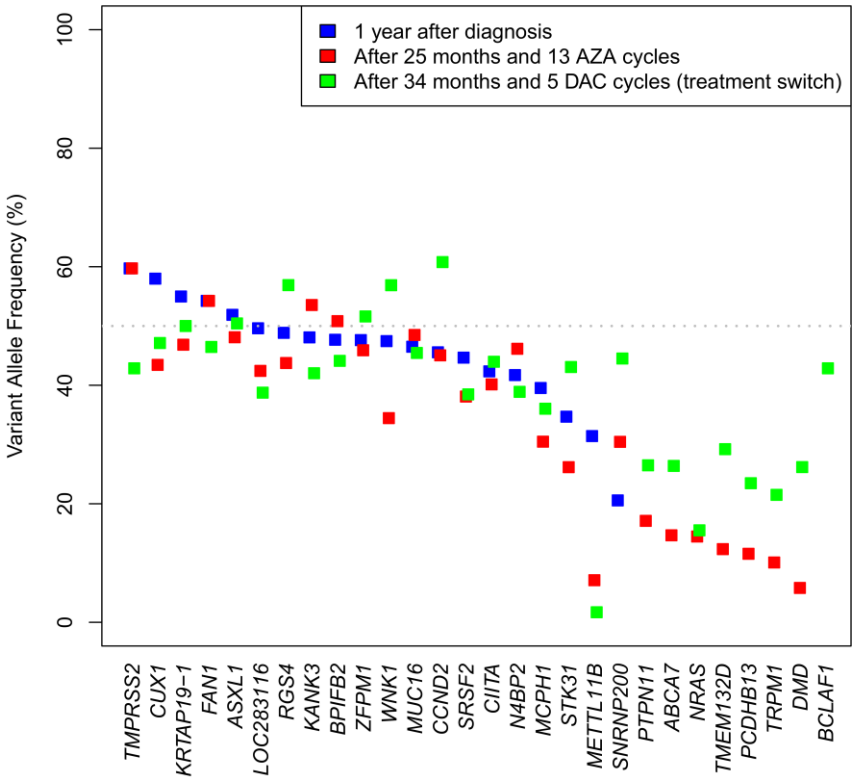
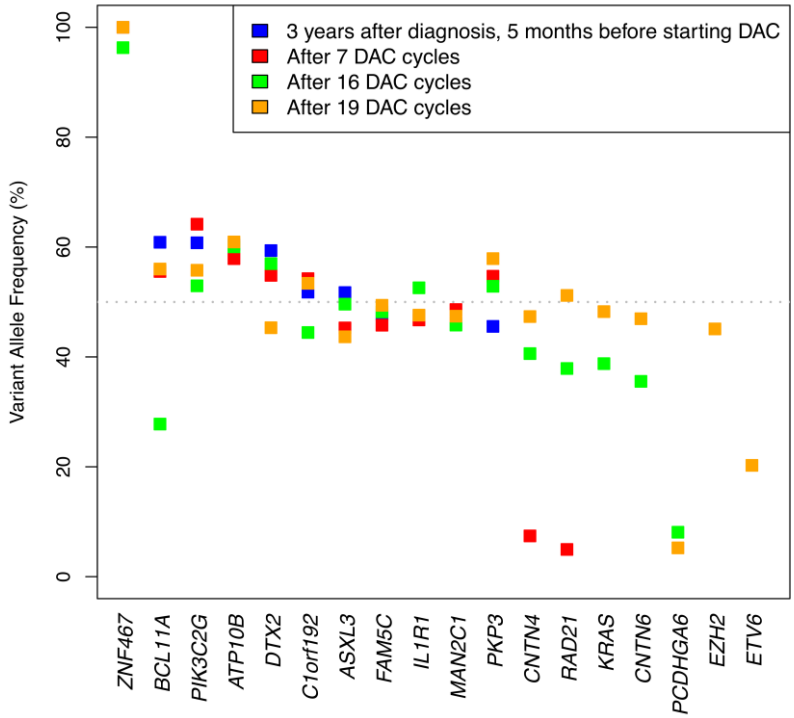
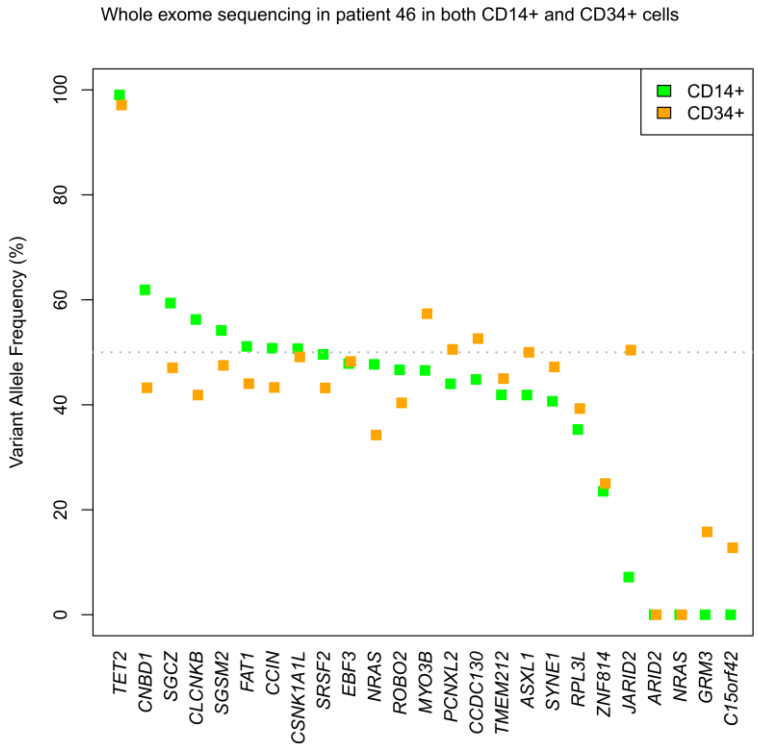


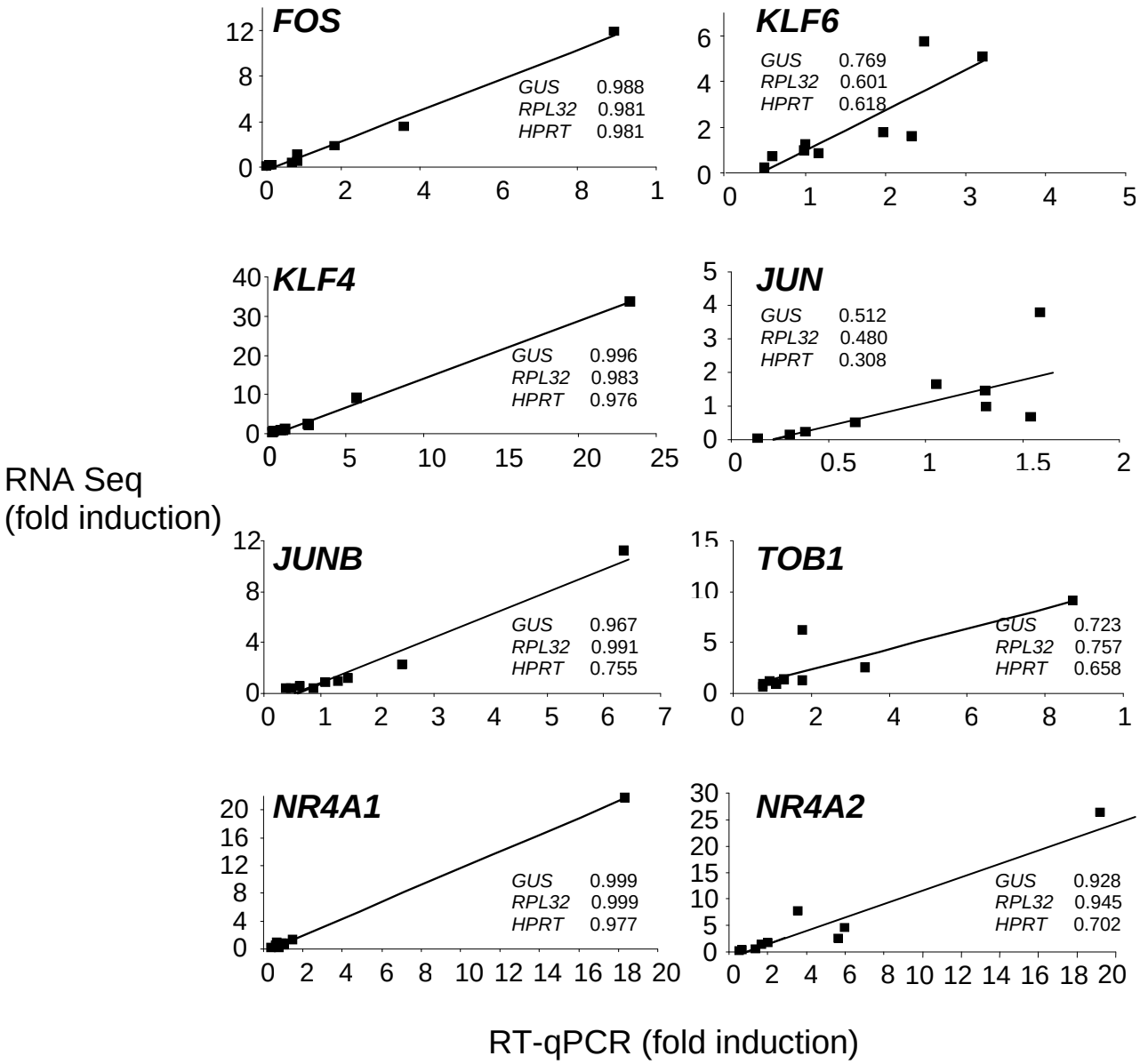
Figure S21: Serial whole exome sequencing in patient 34



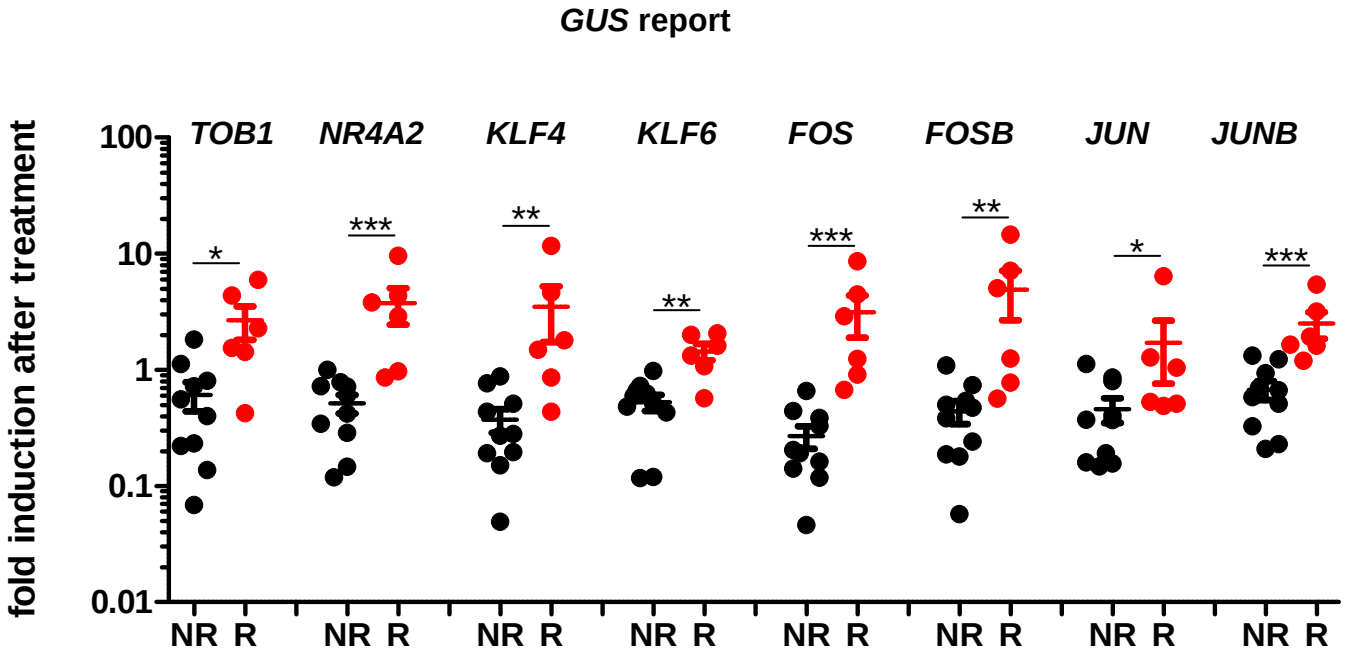
Supplementary Figure 22: Comparison of whole exome sequencing in sorted CD14+ and CD34+ cells for UPN46 after treatment

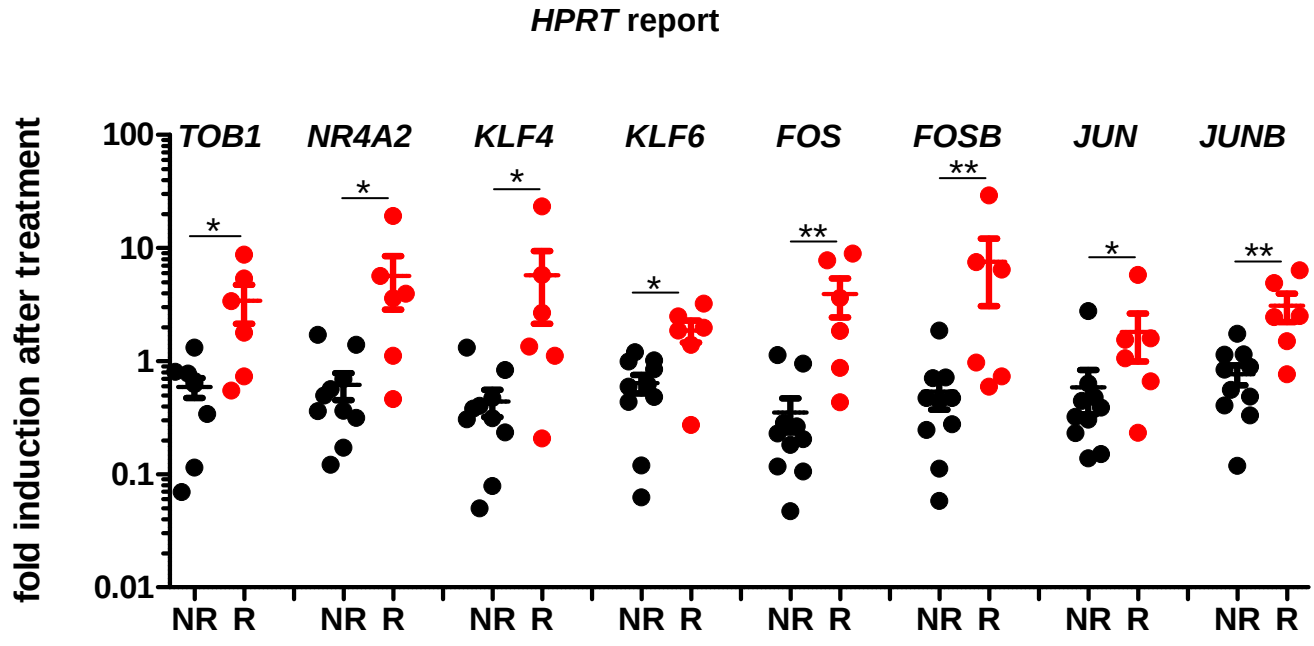


Supplementary Figure 23: RNA-Seq data validation. We compared indicated gene expression measured at two different time points in 9 patients by RNA-Seq and reverse transcription – quantitative polymerase chain reaction (RT-qPCR), respectively. For each studied gene, we measured gene expression induction at the second time-point compared to the first one. Correlations between the two methods are shown. RT-qPCR data were normalized to three independent reporter genes (*GUS*, *RPL32* and *HPRT*). Results plotted are those obtained with *GUS* normalization. R squared values generated by using RT-qPCR data obtained with each reporter are shown.



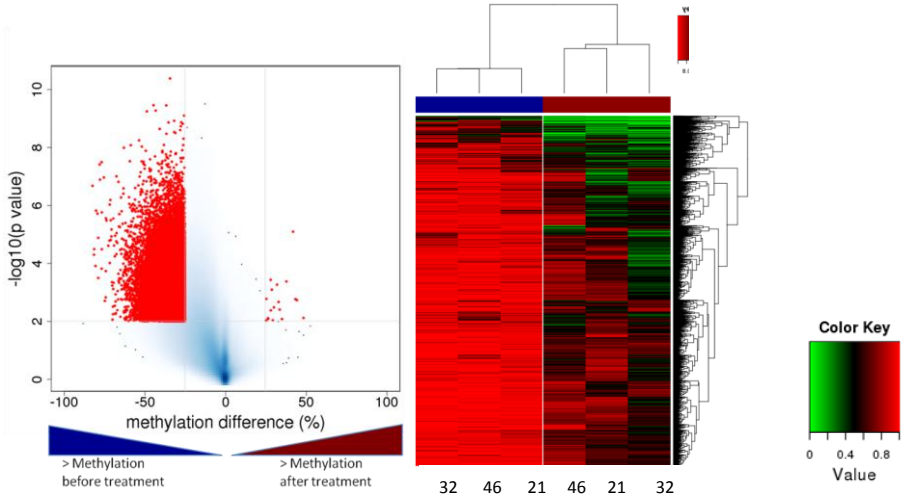
Supplementary Figure 24: Validation of RNA sequencing data. The differential expression of eight genes was explored in 6 responders (R, 3 studied by RNA-Seq in Figure 6b and 3 additional cases) and 10 non-responders (NR, 3 studied by RNA-Seq in Figure 6a and 7 additional cases). Data obtained with 2 housekeeping genes (*GUS* and *HPRT*) complement those obtained by using RPL32 as normalizer



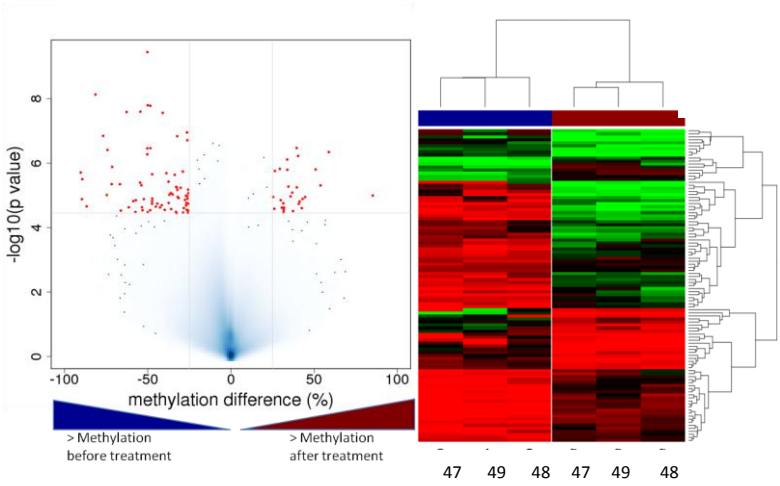


Supplementary Figure 25: Volcano plots and heatmaps of differentially methylated regions. These regions were studied twice in 9 patients, including 6 patients treated with a demethylating agents (responders: 3, non responders: 3) and 3 untreated patients.

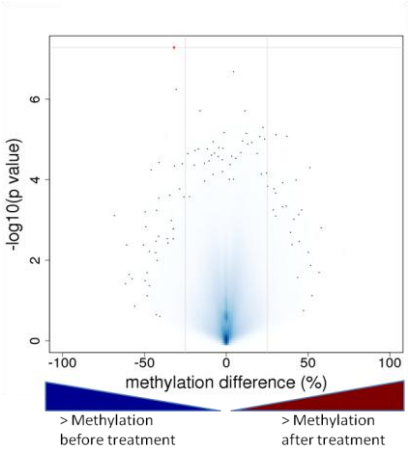
A - Responders



B – Non-responders (stable disease)

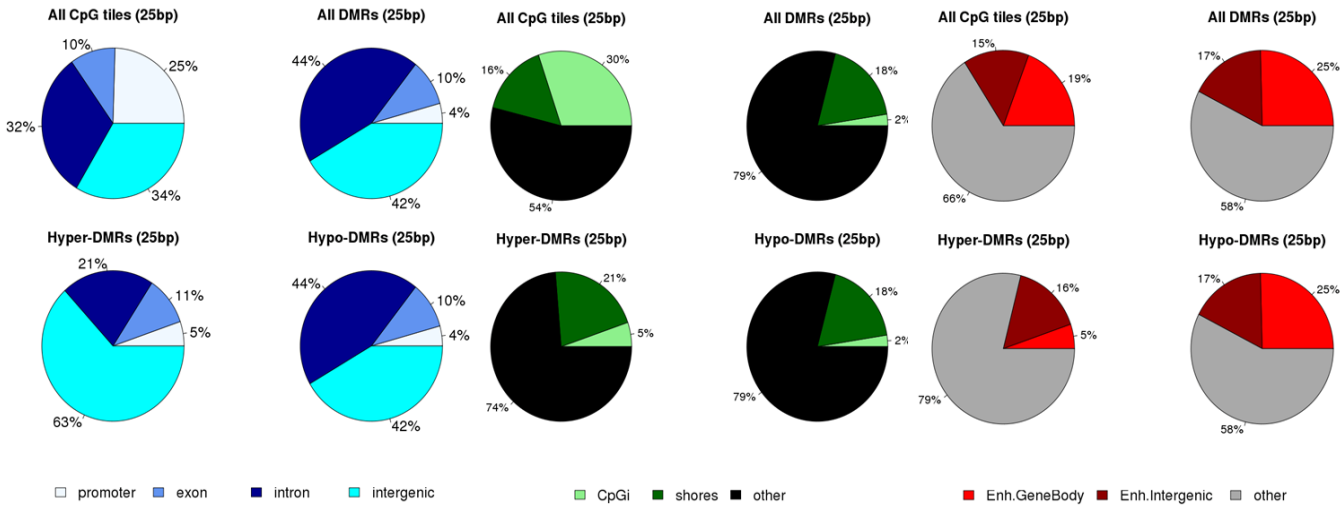


C - Untreated

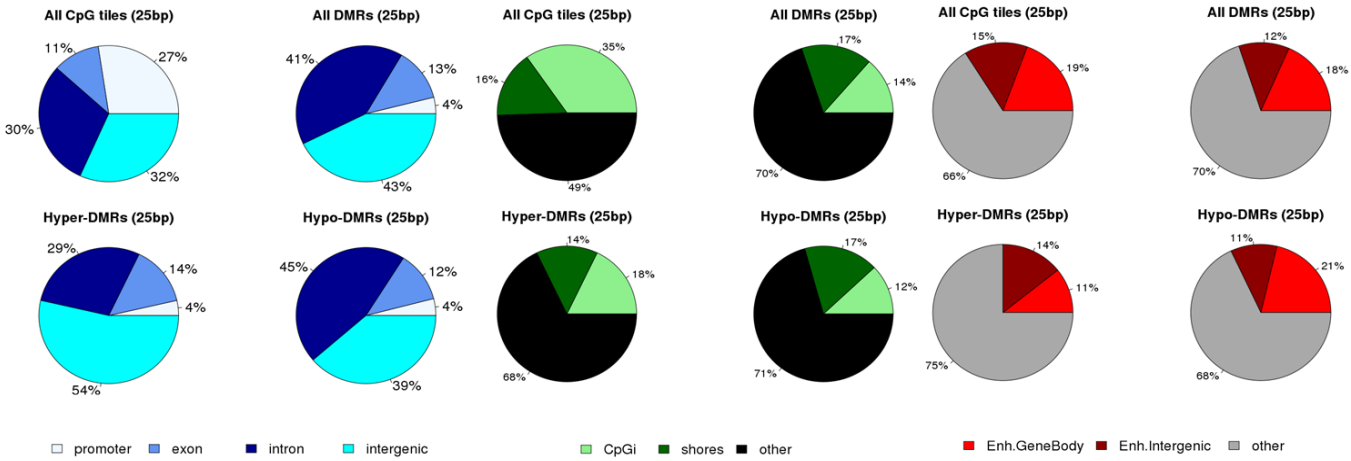


Supplementary Figure 26: Analysis of differentially methylated region (DMR) repartition. Pie charts illustrate the relative proportion of CpG tiles and DMRs annotated to the RefSeq promoter, exonic, intronic and intergenic regions, annotated to CpG islands, CpG shores and regions beyond CpG shores, and finally annotated to enhancers within gene bodies, intergenic and nonenhancer regions. **A**. In responders, differentially methylated regions were significantly depleted in promoters (5% vs. 24%, $p < 2.2\times10^{-16}$), as well as in CpGi (2% vs. 29%, $p < 2.2\times10^{-16}$), and significantly enriched in generic enhancers ($p < 2.2\times10^{-16}$). **B**. In non-responders (stable disease) remaining on therapy, DMRs, which were quantitatively much less important (see Figure 4), were also significantly depleted in promoters (4% vs. 27%, $p < 4.61\times10^{-10}$), as well as in CpGi (14% vs. 35%, $p < 1.14\times10^{-6}$), and hypo-DMRs were significantly enriched in generic enhancers compared with hyper-DMRs (21% vs. 11%, $p < 0.005$).

A - Responders



Non-responders (stable disease)



II - Supplementary Tables

Note: Five large tables are given separately

Supplementary Table 1: Characteristics of the three studied cohorts of patients.

	Whole exome (N=49)	Whole genome (N=17)	Validation (N=180)
Clinical Information			
Age in years : median (range)	74 (45-89)	71 (58-91)	74 (46-93)
Gender, n (%)			
Male	30 (61)	13 (76)	127 (71)
Female	19 (39)	4 (24)	53 (29)
Prior evolution, n (%)			
<6 months	28 (57)	8 (53)	80 (44)
>6 months	21 (43)	7 (47)	100 (56)
NA	0	2	0
WHO diagnosis			
CMML-1	37	15	158
CMML-2	12	1	22
NA	0	1	0
Cytogenetic risk*, n (%)			
Low	36 (74)	11 (65)	135 (75)
Intermediate	7 (14)	1 (6)	27 (15)
High	6 (12)	0	7 (4)
NA		5 (29)	11 (6)
WBC, 10^9/L, median (range)	14.7 (3.3-133,0)	8.8 (3.9-58)	10.3 (2.3-366.8)
Hemoglobin, g/dL, median (range)	11.8 (4.2-15.6)	12.2 (8.1-15.8)	11.3 (4.9-26.8)
Platelets, 10^9/L, median (range)	119 (6-1051)	104 (13-320)	117 (3-1427)
Monocytes, 10^9/L, median (range)	1.9 (1.0-62.8)	1.66 (1.2-17.8)	1.97 (1.0-84.4)
Peripheral Blasts %, median (range)	0 (0-14)	0 (0-6)	0 (0-9)
Bone Marrow Blasts %, median (range)	5 (0-17)	6 (2-18)	4 (0-18)
Immature myeloid cells,%, median (range)	0 (0-31)	0 (0-29)	0 (0-34)
Extramedullary-disease, n (%)			
Present	7 (15)	1 (6)	44 (24)
Absent	33 (67)	13 (76)	131 (73)
NA	9 (18)	3 (18)	5 (3)
Mutational status (%)			
TET2	59	65	64
SRSF2	47	18	44
ASXL1	33	12	32
CBL	20	12	16
KRAS	16	6	14
NRAS	16	12	13
DNMT3A	12	0	6
U2AF1	10	6	11
RUNX1	10	6	20
SF3B1	10	0	10
ZRSR2	8	6	11
CUX1	6	0	ND
EZH2	6	6	9

IDH2	6	0	12
LUC7L2	6	0	0
BCOR	6	0	0/68
JAK2	4	0	12
SH2B3	4	0	ND
ETNK1	4	0	3
NF1	6	0	6
ASXL2	4	0	0
DOCK2	4	0	2
ABCC9	4	0	2
HUWE1	4	0	1
TTN	4	0	ND
PHF6	4	0	8

evolution : time from diagnosis to sampling

* According to the Spanish CMML cytogenetic classification. Low: normal and isolated -Y ; intermediate : other abnormalities and high : trisomy 8, complex caryotypes (>=3 abnormalities) and abnormalities of chromosome 7.

Immature myeloid cells include promyelocytes, myelocytes and metamyelocytes detected in the peripheral blood. NA, not available. ND, not done.

Supplementary Table 2: Whole exome sequencing and gene re-sequencing. Control samples were either sorted $CD3^+$ lymphocytes or skin fibroblasts or buccal swabs. T1, T2, T3, T4, T5 indicate the numbering of serial sequencing.

Whole exome sequencing						
Sample type	**Control**	**Tumor - T1**	**Tumor - T2**	**Tumor - T3**	**Tumor - T4**	**Tumor -T5**
Nb of samples	N=49	N=49	N=17	N=6	N=3	N=1
Total reads (Mean)	135257990	129478657	142904800	105178277	98540375	125196156
%uniqMach (Mean)	94.85	94.74	94.29	93,67	94,4	96,365
Total bases (Mean)	9011607882	8928770091	9511603200	7973397798	7376183179	9724634877
%onTarget (Mean)	62.98	62.99	62.84	64,57	61,23	61,674
1x (Mean)	83.06	83.32	93.47	97,5	96,52	98,817
10x (Mean)	78.49	79.4	88.46	93,66	91,63	98,38
20x (Mean)	73.21	74.7	82.1	90,01	86,56	97,447
Coverage (Mean)	112.31	111.22	122.29	102,8	89	119
Coverage (SD)	47.88	40.46	71	58,85	45,43	NA
Coverage (Range)	20-317	24-231	10-259	30-201	37-121	NA
Gene re-sequencing						
Sample type	**Control**	**Tumor - T1**	**Tumor - T2**	**Tumor - T3**	**Tumor - T4**	
20x (Mean)	93.07	92.59	91.88	90.7	91.41	
Coverage (Mean)	755.32	756.26	785.73	634	861.5	
Coverage (SD)	238.5	255.48	494.21	176.78	754.48	
Coverage (Range)	386-1460	376-1577	314-1984	509-759	328-1395	

NA: not applicable

Supplementary Table 3: List of the 680 somatic mutations validated by re-sequencing

See independent xls file.

Supplementary Table 4: Targeted re-sequencing of recently identified and previously unknown recurrently mutated genes. (*ASXL2*, *PHF6*, *DOCK2*, *NF1*, *ABCC9*, *HUWE1*, *ETNK1*, *LUC7L2*). Mutations were validated using MiSEq in 180 samples.

Sample type	**Tumor**
Nb of samples	N=180
20x (Mean)	92.02
Coverage (Mean)	690.11
Coverage (SD)	459.62
Coverage (Range)	96-2888

Supplementary Table 5: List of variants detected by targeted re-sequencing of previously unknown recurrently mutated genes.

Gene	Mutation type	RefSeq	Amino acid change	Nucleotide change	Mutated patients (training set N=49)	Mutated patients (validation set N=180)
PHF6	Nonsynonymous	NM_001015877	I314T	T941C	1	4
	FDel	NM_001015877	K26fs	76delA	0	1
	Stopgain	NM_001015877	E27X	G79T	0	1
	Stopgain	NM_001015877	L31X	T92A	0	1
	FDel	NM_001015877	G186fs	559delG	0	1
	Stopgain	NM_001015877	R225X	C673T	0	1
	FDel	NM_001015877	M243fs	729delG	0	1
	Nonsynonymous	NM_001015877	V268A	T803C	0	1
	Nonsynonymous	NM_001015877	R274Q	G821A	0	1
	Nonsynonymous	NM_001015877	G287D	G860A	1	0
	Nonsynonymous	NM_001015877	A288T	G862A	0	1
	Stopgain	NM_001015877	R319X	C955T	0	1
	Splice	NM_001015877		730-1G>T	0	1
	Splice	NM_001015877		1098+1G>A	0	1
NF1	FDel	NM_000267	260_260del	779_780delCC	0	1
	Nonsynonymous	NM_000267	V288M	G862A	0	1
	FInsert	NM_000267	*P370fs	1108_1109insCC	0	1
	FInsert	NM_000267	T676fs	2027_2028insC	0	1
	Nonsynonymous	NM_000267	L792H	T2375A	0	1
	Nonsynonymous	NM_000267	N793T	A2378C	0	1
	Nonsynonymous	NM_000267	R1276Q	G3827A	1	0
	Nonsynonymous	NM_000267	Y1587C		0	1
	Nonsynonymous	NM_000267	L1339R		0	1
	Stopgain	NM_000267	R1748X	C5242T	0	1
	Nonsynonymous	NM_000267	S1997N	G5990A	0	1
	Nonsynonymous	NM_000267	R2237Q	G6710A	0	1
	Splice	NM_000267		1185+1G>C	0	1
	Splice	NM_000267		A4760G	1	0
	Splice	NM_000267		204_205-2delAG	1	0
DOCK2	Nonsynonymous	NM_004946	M770V	A2308G	0	1
	Nonsynonymous	NM_004946	C853F	G2558T	0	1
	Nonsynonymous	NM_004946	R1189W	C3565T	1	0
	Nonsynonymous	NM_004946	L1208V	C3622G	1	0
	Splice	NM_004946		1258+5G>C	0	1
ABCC9	FDel	NM_005691	F472fs	1416delT	0	1
	Nonsynonymous	NM_005691	E607G	A1820G	1	0
	Nonsynonymous	NM_005691	T621I	C1862T	1	0
	FDel	NM_005691	*D1439fs	4317delT	0	1
	Splice	NM_005691	*	3566+1G>A	0	1
HUWE1	Nonsynonymous	NM_031407	R629H	G1886A	1	0
	Nonsynonymous	NM_031407	A4058V	C12173T	1	0
	Splice	NM_031407		2261+9T>G	0	1
ASXL2	Stopgain	NM_018263	R614X	C1840T	1	0
	FDel	NM_018263	E1172fs	3515delA	0	1

PHF6: 18 abnormalities were identified in 17 patients, including 14 distinct abnormalities and 2 variants in one of the patients (I314T and E27X); *NF1*: 15 abnormalities were identified in 14 patients, one patient carrying two variants, T676fs and L1339R; *ASXL2*: in addition to the shown variants, we detected 3 potentially germline SNVs (A497T, S185G, Q1371K) in 6, 3 and 2 patients respectively; smaller frequencies being reported in public databases; * indicates potential germline variants with an allelic frequency of 50 or 100% and no information in 1000G or ESP.

Supplementary Table 6: Correlations between mutated genes and clinical and biological parameters. We used Fisher exact test for qualitative parameters and Wilcoxon tests for quantitative parameters.

	Positive correlation	Negative correlation
Hemoglobin level	*TET2 (***)*	*ASXL1 (**), SF3B1 (**), ZRSR2 (**)*
Platelet count	*SF3B1 (***)*	*RUNX1 (***), SRSF2 (***)*
White blood cell count	*JAK2 (**), NRAS (***), ASXL1 (***)*	*PHF6 (***)*
Monocyte count	*NRAS (***), ASXL1 (**), SRSF2 (**)*	
Peripheral blast cell count	*ASXL1 (***), LUC7L2 (**)*	*TET2 (***)*
Immature myeloid cell count	*ASXL1 (***)*	
Medullary blast percentage	*KRAS (**)*	*JAK2 (***)*
CMML2 WHO subgroup	*KRAS (**)*	
Low cytogenetic risk		*TET2 (**)*

**: 0.001<P≤0.01 ; ***: P≤0.001

Supplementary Table 7: Alignment and coverage of whole genome sequencing

Sample type	Control	Tumor
Nb of samples	N=17	N=17
Total reads (Mean)	1028029614	1084281186
%uniqMach (Mean)	99.46	99.38
Total bases (Mean)	86842279017	91918801269
Nb covered bases (Mean)	2820703699	2819926295
Coverage (Mean)	30.18	32.12
Coverage (SD)	4.3	9.9
Coverage (Range)	27-44	25-59

Supplementary Table 8: Somatic variants (N=8077) detected by whole genome sequencing

See independent xls file.

Supplementary Table 9: Somatic variants identified in hotspot (N=46), promoter (N=147) and enhancer (N=37) regions of the genome. The three enhancers predicted to be active in hematopoietic cells are in bold in this table.

See independent xls file.

Supplementary Table 10: Time between consecutive genomic analyses. The mean time between the two first time points in whole exome analyses was 12+/- 8 months in responders, 12 +/- 7 in non responders (stable disease), and 13 +/- 9 months in untreated patients.

UPN	Treatment Response	T1-T2 (months)	T2-T3 (months)	T3-T4 (months)	T4-T5 (months)	Mean (months)
1	Responder	21				21
3	Stable disease	12				12
5	Untreated	13				13
9	Untreated	28				28
21	Responder	7	18			12,5
23	Untreated	16				16
28	Stable disease	25,5	9			17,25
29	Untreated	4				4
30	Untreated	4,5				4,5
32	Responder	1	11	5	20	9,25
33	Untreated	15				15
34	Responder	12	9	3		8
35	Stable disease	8				8
46	Responder	17	47			32
47	Stable disease	6	26	3		11,7
48	Stable disease	9				9
49	Stable disease	12				12
Mean		**12,4**	**20**	**3,7**	**20**	**13,7**
SD		7,5	14,8	1,1	NA	7,5
Range		1-28	9-47	3-5	NA	4-32

Supplementary Table 11: Samples used for serial RNA-Seq and methylation experiments. AZA, azacytidine; DAC, Decitabine

UPN	Treatment	Status	RNA-Seq	ERRBS
UPN5	No		Yes	Yes
	No		Yes	Yes
UPN23	No		Yes	Yes
	No		Yes	Yes
UPN9	No		Yes	Yes
	No		Yes	Yes
UPN21	No		Yes	Yes
	AZA	Responder	Yes	Yes
UPN32	No		Yes	Yes
	AZA	Responder	Yes	Yes
UPN46	No		Yes	Yes
	DAC	Responder	Yes	Yes
UPN47	No		Yes	Yes
	DAC	Non Responder	Yes	Yes
UPN48	No		Yes	Yes
	DAC	Non Responder	Yes	Yes
UPN49	No		Yes	Yes
	DAC	Non Responder	Yes	Yes

Supplementary Table 12: RNA sequencing data alignment and coverage

Sample	RawR1/2	FilteredR1	FilteredR2	Left reads mapped	Right reads mapped	Over. Read alignment rate	Aligned pairs	concordant pair alignment rate	Reads On Transcriptome
UPN32_808	46808962	41294065	33925275	32364485 (78.4%)	30208145 (89.0%)	83.2%	28057834	79.7%	18060508
UPN32_1054	71231589	63156201	51852212	51366562 (81.3%)	47256823 (91.1%)	85.8%	44301062	82.2%	38728462
UPN46_227	120224282	117345119	110433131	109556388 (93.4%)	105364524 (95.4%)	94.4%	103453840	92.7%	88272219
UPN46_936	115308707	111565858	100280322	102577193 (91.9%)	94020431 (93.8%)	92.8%	92194563	90.7%	78459035
UPN21_969	122942488	119596084	108717931	110169667 (92.1%)	101511755 (93.4%)	92.7%	99702165	90.4%	88226310
UPN21_1284	52230173	48868690	32924089	40968346 (83.8%)	27634462 (83.9%)	83.9%	26937503	79.7%	34416112
UPN47_320	49482782	46217919	31099711	38350179 (83.0%)	25936814 (83.4%)	83.1%	25234168	79.0%	31235209
UPN47_408	108286314	104859767	94984982	95921983 (91.5%)	89302431 (94.0%)	92.7%	87370953	90.8%	76465420
UPN48_299	121708305	117891046	107449531	108434319 (92.0%)	101166905 (94.2%)	93.0%	99076424	90.9%	89400337
UPN48_426	38691148	32353829	16501191	22918394 (70.8%)	14644836 (88.8%)	76.9%	13074220	76.8%	2263141
UPN49_257	34995060	31286727	15008212	23192812 (74.1%)	13540416 (90.2%)	79.3%	12073219	75.1%	17626652
UPN49_433	118627675	116466600	108720777	106239580 (91.2%)	102765772 (94.5%)	92.8%	100519319	91.7%	83054517
UPN23_704	85241114	83736952	78455347	76106552 (90.9%)	73854796 (94.1%)	92.5%	72231454	91.4%	63159618
UPN23_939	133832338	130785724	120054128	116728703 (89.3%)	111186169 (92.6%)	90.9%	108613765	89.7%	93059490
UPN9_608	71956594	70176553	64136555	62891114 (89.6%)	59822073 (93.3%)	91.4%	58393627	90.2%	52557797
UPN9_1009	118307267	116598822	108180366	105229340 (90.2%)	102000602 (94.3%)	92.2%	99604907	91.5%	83850048
UPN5_732	45571594	44875674	42188768	40633204 (90.5%)	39893674 (94.6%)	92.5%	38963966	91.8%	33764802
UPN5_870	57207766	55328939	51138722	48906249 (88.4%)	47941566 (93.7%)	91.0%	46278280	89.4%	39243634

Supplementary Table 13: Effect of time on gene expression. List of differentially expressed genes (abs(log2FoldChange)>=1) between two time points are given for 3 responders (N=513) and 3 non responders / stable disease (N=63), by separating up- and down-regulated genes in each category. No change was observed in the three studied untreated patients (see Table 1). **A**. Non responders (stable disease); **B**. Responders

See independent xls file.

Supplementary Table 14: Effect of time on genome methylation. List of differentially methylated regions (having>=25% difference) in three responders (N=35,914) and three non responders / stable disease (N=103) between 2 sampling is provided by separating up and down-methylated regions. No change was observed in the three studied untreated patients.

See independent xls file.

Altérations d'expression génique dans la leucémie myélomonocytaire chronique

6

Des mutations somatiques (14 en moyenne) sont constamment retrouvées dans les cellules de patients atteints de *LMMC*. Ces mutations sont-elles responsables de la maladie et de son expression clinico-biologique ? La découverte de mutations de gènes tels que *TET2*, *DNMT3A*, *ASXL1*, *JAK2* dans les cellules sanguines circulantes de sujets âgés sans déséquilibre de l'hémogramme ni symptôme clinique suggère que la présence de mutations ne suffit pas à induire la maladie (Busque et al. (2012), Xie et al. (2014), Jaiswal et al. (2014), Genovese et al. (2014)). La découverte récente de clones (jusqu'à 6 par mm^2 de peau saine) dans lesquels plusieurs des mutations somatiques couramment identifiées dans les carcinomes cutanés pouvaient être associées (Martincorena et al. (2015)), montre que la présence de mutations somatiques dans des gènes clefs ne suffit pas pour provoquer une tumeur. De nombreux autres paramètres doivent intervenir, parmi lesquels le microenvironnement et la réponse immunitaire. Dans le contexte des hémopathies myéloïdes, ces rôles sont encore très mal cernés.

Un autre aspect de la maladie reste confus : il s'agit du lien entre anomalies génétiques et épigénétiques. Qu'est ce qui déclenche la maladie ? S'agit-il des modifications épigénétiques associées à l'âge qui favorisent l'accumulation de mutations ? Ou ces modifications sont-elles la conséquence de l'accumulation de mutations dans des gènes régulateurs de l'épigénétique ? Ces mêmes questions pourraient s'appliquer à l'épissage des ARN pré-messagers.

Nous ne répondons pas à ces questions. Nous nous sommes limités à étudier les anomalies de l'expression des gènes et de l'épissage des ARN pré-messagers dans les monocytes de patients atteints de leucémie myélomonocytaire chronique. Les résultats ci-dessous concernent le séquençage d'*ARN* ribodéplétés des monocytes de 10 patients *LMMC* (6 patients mutés $SRSF2^{P95}$ et 4 patients non mutés pour *SRSF2*, *U2AF1*, *SF3B1* et *ZRSR2*) et 4 sujets contrôles d'âge similaire. Les échantillons ont été séquencés de façon à avoir un nombre de lectures suffisant pour faire une quantification de l'expression génique. La profondeur atteinte pour deux échantillons (1 patient et 1 contrôle) s'est avérée insuffisante (tableau A.1 p174).

6.1 Gènes anormalement exprimés

Après avoir quantifié le niveau d'expression de chaque gène dans chaque échantillon comme décrit dans la partie 4.2, nous avons procédé à la normalisation des données (figure A.4 p174). L'analyse d'expression différentielle a été réalisée avec le package DESeq2 et la version 3.2.1 de R. Les outliers ont été remplacés par la moyenne de l'expression de tous les échantillons excepté lui-même. Avec la procédure independentFiltering proposée dans DESeq2, nous avons éliminé les gènes dont l'expression est si basse qu'ils n'ont que très peu de chances d'être détectés comme

différentiellement exprimés. De cette manière, avec la suppression de 37.5% des données, le nombre de rejets de l'hypothèse nulle ("le gène n'est pas différentiellement exprimé") a atteint son maximum (figure A.5 p175). Cette procédure nous a permis de détecter 198 gènes différentiellement exprimés supplémentaires. Nous avons ainsi observé la dérégulation de 1623 gènes d'au moins un facteur 2, avec 1095 gènes sur-exprimés et 528 gènes sous-exprimés chez les patients. Nous avons ainsi retrouvé la dérégulation fréquente de CJUN déjà signalée par notre équipe.

La figure 6.1 indique :
— les 1623 gènes avec une padj$\leq$ 0.01 et $|log2FoldChange| \leq 1$ en rouge
— les 3954 gènes avec $|log2FoldChange| \leq 1$ en orange
— les 2106 gènes avec une padj$\leq$ 0.01 en vert
— tous les autres gènes en noir

Pour plus de lisibilité, seuls les 18 gènes les plus significatifs (padj$\leq$ 10^{-10} et $|log2FoldChange| \leq 2$) ont été étiquetés. Le nombre important de gènes dérégulés souligne le caractère dysplasique de la pathologie. Les cinq gènes les plus significativement sur-exprimés sont CDKN2D, ID1, ARL4A, MALAT1 et SH3D19. Une sur-expression de ID1 a été observée chez les patients atteints de leucémie aiguë *de novo* (Zhou et al. (2015)). Une sur-expression de MALAT1 (metastasis associated lung adenocarcinoma transcript 1), un long *ARN* non codant qui promeut la croissance tumorale en régulant le cycle cellulaire, a été décrite dans de nombreux cancers solides (Zhang et al. (2015)) : cancer du rein (Hirata et al. (2015)), gliome (Ma et al. (2015)), cancer de l'œsophage (Hu et al. (2015)). Des travaux en cours dans l'équipe, en collaboration avec celle d'Eric Padron à Tampa (Floride), suggèrent la perturbation d'un lien fonctionnel entre MALAT1 et SRSF2. Les cinq gènes les plus significativement sous-exprimés sont IL32, PARS2, ZNF2, ABAT et BAIAP2. Jansen et al. (2015) ont identifié ABAT comme biomarqueur prédictif de la résistance hormonothérapie dans le cancer du sein.

Si l'on réalise une étude des gènes par famille, on constate une sur-expression chez les patients de gènes des familles :
— Nuclear Factor kappa binding (NFκB) : NFκBIA, NFκBIZ, NFκBID et NFκBIE
— Kruppel-like factor (KLF) : KLF9, KLF6, KLF10, KLF7 et KLF4
— Cyclin-dependent kinase inhibitor (CDKN) : CDKN2D ou p19, CDKN1A ou p21, CDKN3 et CDKN2C ou p18
— Proto-oncogene, serine/threonine kinase (PIM) : PIM3 et PIM1
— Inhibitor of DNA binding (ID) : ID1 et ID2
— Famille apoptotique : MCL1 et BCL2A1.

Une analyse de pathways réalisée avec DAVID (Database for Annotation, Visualization and Integrated Discovery) est présentée dans le tableau 6.1. Nous avons considéré comme significatifs les pathways avec un taux de faux positif (FDR, False Discovery Rate) inférieur à 0.05. Les analyses de pathway servent à titre indicatif, bien souvent la majorité des gènes n'est pas connue dans la base d'annotation et l'analyse ne se base que sur une partie des données. Dans notre cas, seuls 29.6% et 25.1% des gènes dérégulés étaient connus dans KEGG (Kyoto Encyclopedia of Genes and Genomes) pour les gènes sur-exprimés et les gènes sous-exprimés respectivement. Nous observons une dérégulation significative de six voies, pour lesquelles nous indiquons le nombre de gènes altérés ainsi que le *FDR* associé :
— Signalisation des récepteurs T : notamment de NFAT5, TNFα et TNFαIP3
— Signalisation de p53, majoritairement par des sur-expressions de MDM4, P21, GADD45A et CDK6
— Cycle cellulaire : 21 molécules, dont ARFIP1, P53, Gadd45, MAPK6, CDK6, E2F2 et DBF4

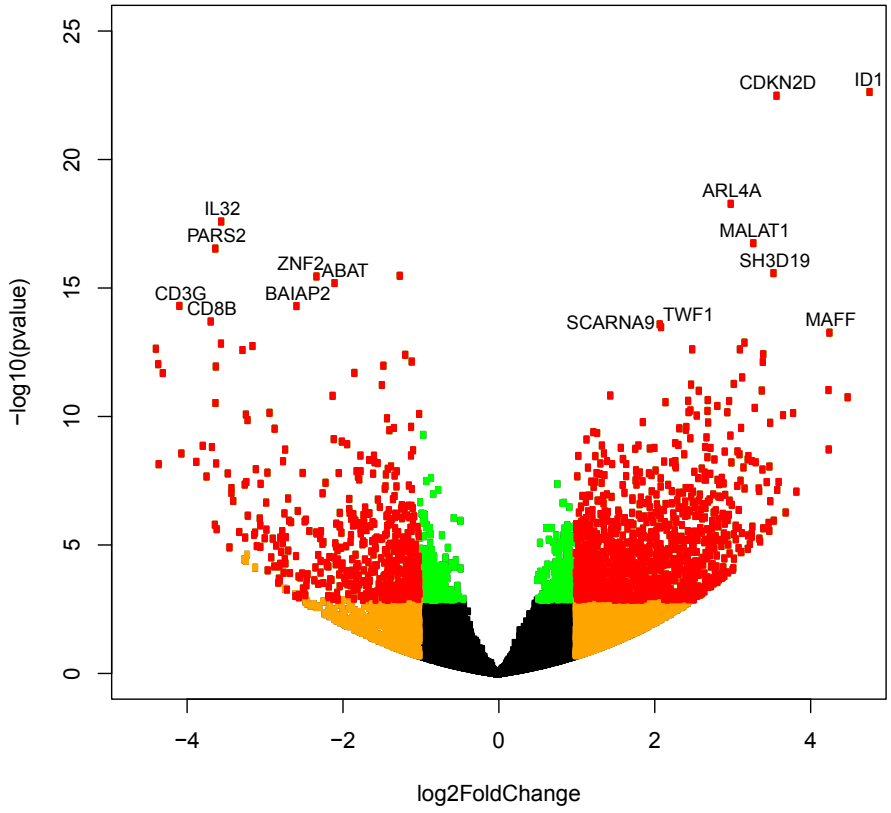


FIGURE 6.1 – *Volcano plot de la comparaison de l'expression des monocytes chez patients LMMC et sujets sains*

	Gènes dérégulés		Sur exprimés chez patients		Sous exprimés chez patients	
Pathways	Nombre	*FDR*	Nombre	*FDR*	Nombre	*FDR*
Signalisation des récepteurs T	22	$5.5.10^{-3}$			12	$2.7.10^{-3}$
Signalisation de p53	15	$1.5.10^{-2}$	14	$1.3.10^{-3}$		
Cycle cellulaire	22	$1.6.10^{-2}$	21	$4.7.10^{-4}$		
Ribosome	17	$1.8.10^{-2}$	16	$1.7.10^{-3}$		
Réponse immunitaire					8	$1.8.10^{-3}$
Lignage hématopoïétique					10	$7.4.10^{-3}$

TABLE 6.1 – *Pathways dérégulés en expression dans la leucémie myélomonocytaire chronique*

— Ribosome : majoritairement par sur-expression, de RPL27, RPL27A, RPL24, RPL31, RPL22, RPL22L1, RPL27, RPL27A, MRPL44, RPS10, RPS16, RPS16P5, RPS21 par exemple
— Réponse immunitaire : sous-expression d'IL7R, LCK, ZAP70, CD3D, CD3E et CD8A, CD8B
— Lignage hématopoïétique : globalement réprimé, avec la sous-expression de IL6R, IL2Rα, IL27R, CD5, CD8A, CD8B, et CD3D, CD3E, CD3G.

L'analyse en composantes principales (ACP) (figure 6.2A) des données d'expression des gènes des 14 individus sépare nettement sujets contrôles et sujets malades par l'axe des abscisses. L'axe des ordonnées en revanche n'explique que 13 % de la variance et ne sépare pas ces échantillons. La mesure de similarité entre les échantillons en figure 6.2B montre qu'une partie des patients est plus similaire aux sujets contrôles qu'aux autres patients et que deux patients sont complètement différents des autres échantillons. Une analyse non supervisée réalisée sur les 600 gènes les plus différentiellement exprimés (figure 6.2C) indique également qu'une partie des patients est plus proche des contrôles que des autres patients. Ceci avait déjà été observé par Braun et al. (2011). En revanche, une analyse non supervisée réalisée sur 500 (figure A.6) ou quelques dizaines (ici 20) de gènes les plus différentiellement exprimés (figure 6.2D) permet une disjonction entre patients et contrôles. Ceci suggère que le niveau d'expression de quelques gènes pourrait suffire à distinguer les monocytes des leucémies myélomonocytaires chroniques des monocytes normaux. Cette signature devra être extraite de nos données puis validée sur une nouvelle cohorte.

6.2 Gènes anormalement épissés

Nous avons recherché les anomalies d'épissage entre sujets contrôles et patients *LMMC*. Nous avons sélectionné les événements assez forts, avec padj ≤ 0.01 et $\Delta\phi \geq 20\%$. L'analyse a été réalisée par Émilie Chautard. Avec ces contraintes, nous avons détecté 674 gènes présentant au moins une altération. Plus précisément, nous avons identifié :

— 593 sauts d'exon,
— 194 sauts de plusieurs exons
— 8 exons mutuellement exclusifs
— 4 sites accepteurs alternatifs
— 6 sites donneurs alternatifs.

Nous avons sélectionné 17 événements détectés entre sujets contrôles et sujets malades, que nous avons testés par Q-*PCR* (tableau A.2 p176). Seize de ces événements ont été confirmés. L'analyse de pathways n'a pas montré de voie significativement dérégulée. Les événements d'épissage se produisent-ils de manière aléatoire suite à un dérèglement de la machinerie d'épissage ?

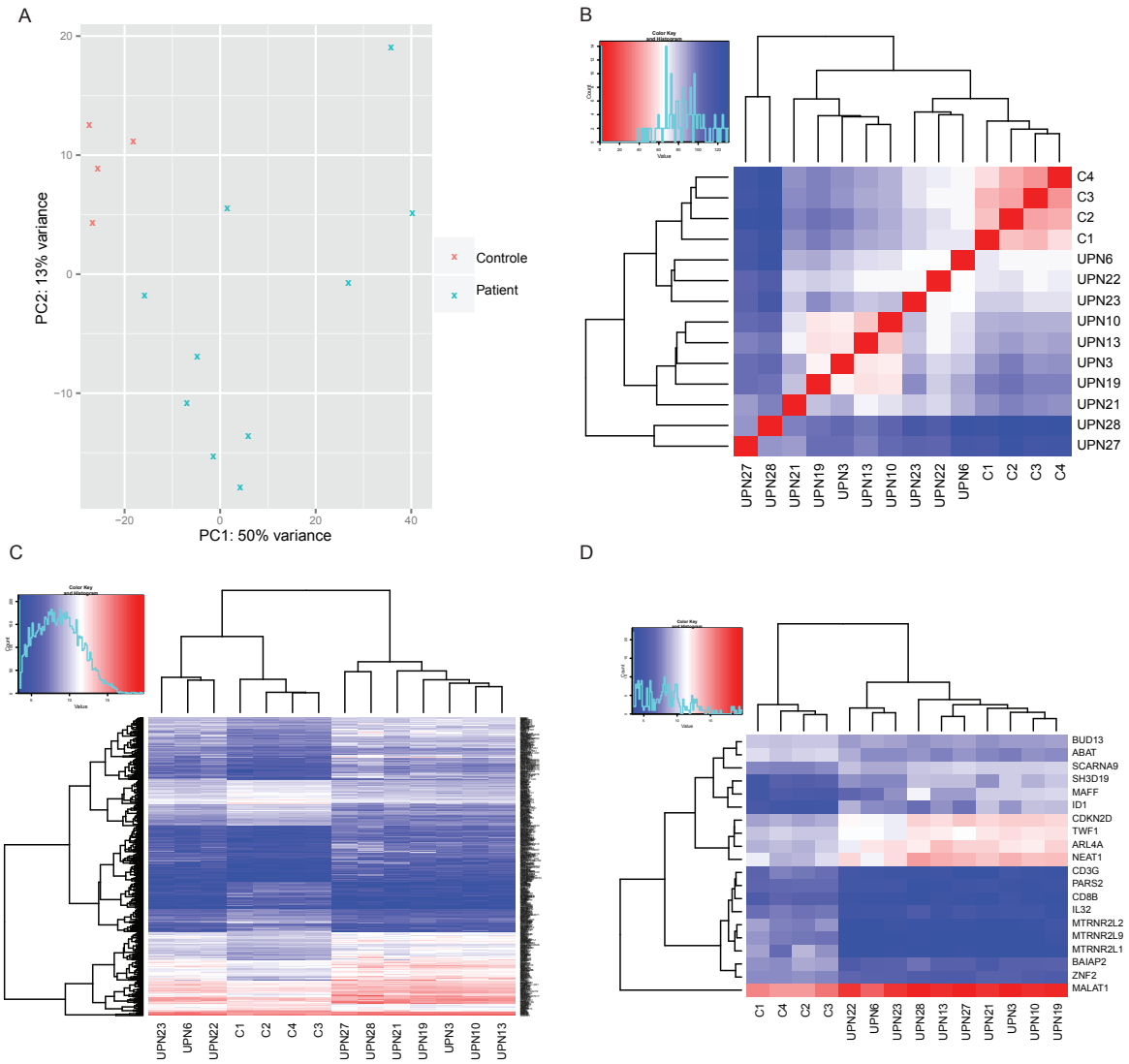


FIGURE 6.2 – *Clusterisation des échantillons de patients et contrôles. A : ACP, B : Matrice de similarité, C : Heatmap des 600 gènes les plus différentiellement exprimés, D : Heatmap des 20 gènes les plus différentiellement exprimés*

FIGURE 6.3 – *Diagrammes de Venn des dérégulations géniques et événements d'épissage alternatif*

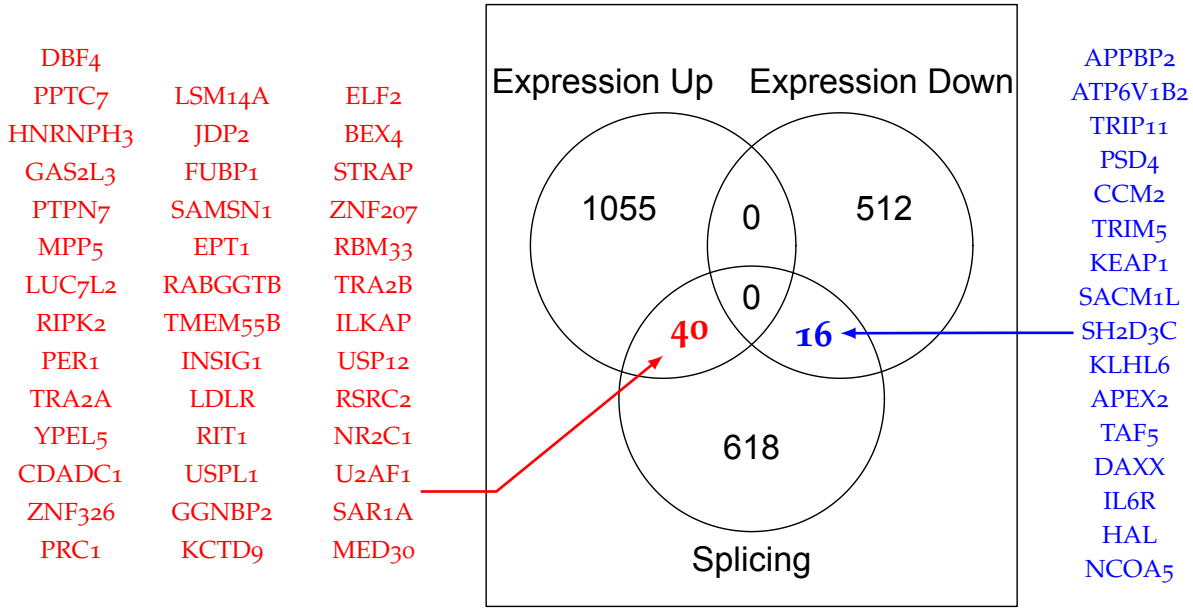


FIGURE 6.4 – *Gènes à la fois anormalement exprimés et anormalement épissés chez les patients*

6.3 Comparaison des gènes anormalement exprimés et anormalement épissés

Puisqu'expression et épissage sont étroitement liés, par exemple du fait du processus NMD (Nonsense-Mediated mRNA Decay), nous avons recherché les gènes ayant subi cette double dérégulation. Pour cela, nous avons comparé les gènes ayant une expression dérégulée (sur ou sous-expression) et les gènes ayant un événement d'épissage alternatif, quelque soit son type. Nous avons identifié seulement 56 événements communs (figure 6.3), soit 8.3% des épissages alternatifs : 16 des gènes sous-exprimés chez les patients et 40 des gènes sur-exprimés chez les patients (figure 6.3). Parmi les derniers, on note en particulier LUC7L2, RIT1 et U2AF1 (figure 6.4), qui sont mutés à une faible fréquence dans la leucémie myélomonocytaire chronique.

L'analyse des pathways réalisée sur les gènes ayant soit des dérégulations géniques soit des anomalies d'épissage (tableau 6.2) indique les pathways dérégulés avec un $FDR \leq 0.05$. Seuls 3 pathways sont significatifs. L'analyse ne montre pas de nouveaux pathways dérégulés, ni ne renforce un des pathways, par rapport aux pathways significativement dérégulés en expression (tableau 6.1) déjà identifiés. Ceci rejoint la question précédente : les événements d'épissage se produisent-ils de manière aléatoire suite à un dérèglement de la machinerie d'épissage ?

	Dérégulations et épissages alternatifs	
Pathways	Nombre	FDR
Signalisation des récepteurs T	29	$1.4.10^{-3}$
Signalisation de p53	21	$1.9.10^{-3}$
Cycle cellulaire	27	$4.6.10^{-2}$

TABLE 6.2 – *Voies modulées par des dérégulations géniques ou par des épissages alternatifs*

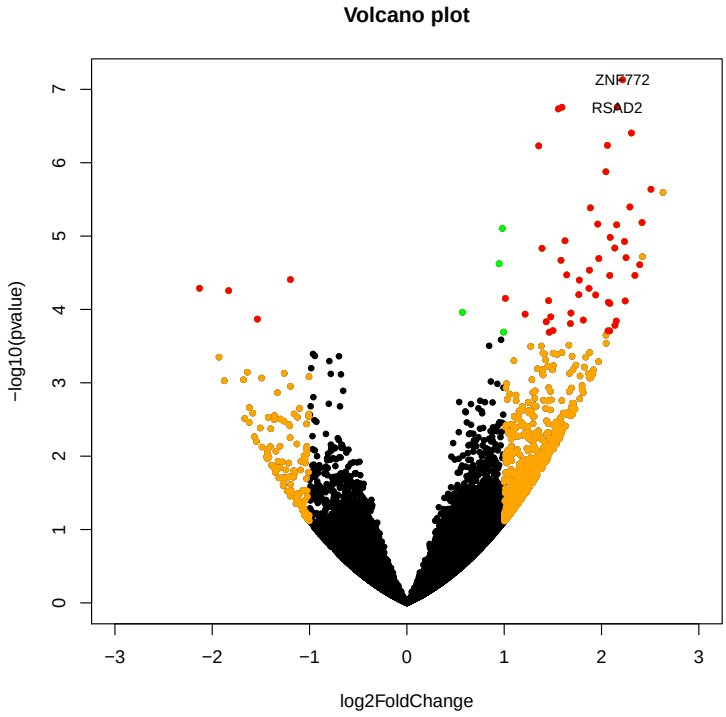


FIGURE 6.5 – *Volcano plot de la comparaison de l'expression des monocytes de patients mutés ou non pour SRSF2. En rouge les 52 gènes avec une padj≤ 0.05 et |log2FoldChange| ≤1. En orange sont les 829 gènes avec |log2FoldChange| ≤1 et en vert les 56 gènes avec une padj≤ 0.05. Pour plus de lisibilité, seuls les 2 gènes les plus significatifs (padj≤ 10^{-3} et |log2FoldChange| ≤2) ont été mentionnés*

6.4 Effet de la mutation de srsf2^{P95}

Nous avons renouvelé l'analyse d'expression différentielle avec le package DESeq2. Nous avons comparé les six patients porteurs de la mutation *SRSF2*^{P95} (UPN3, 6, 22, 23, 27 et 28) aux quatre patients non mutés pour *SRSF2*, *U2AF1*, *SF3B1* et *ZRSR2* (UPN10, 13, 19 et 21). La recherche du pourcentage de gènes à éliminer par la procédure independentFiltering proposée dans DESeq2 a conduit à se débarrasser de 46.6% des gènes. De cette manière, le nombre de rejets de l'hypothèse nulle atteint son maximum (figure A.7). Cette procédure ne nous a pas permis de détecter de gènes différentiels supplémentaires. Nous avons observé la dérégulation de 52 gènes d'au moins un facteur 2 (padj≤0.05), avec 48 gènes sur-exprimés et 4 gènes sous-exprimés chez les patients mutés. Ce faible nombre de dérégulations, visible sur le Volcano plot figure 6.5, suggère que la mutation de *SRSF2* n'a qu'un effet très limité sur l'expression génique dans les monocytes de patients.

L'*ACP* (figure 6.6A) réalisée sur les données d'expression de l'ensemble des gènes des 10 patients ne montre pas de séparation entre patients mutés et patients non mutés bien que 75% de la variance soit expliquée par les deux premiers axes. La mesure de similarité en figure 6.6B entre

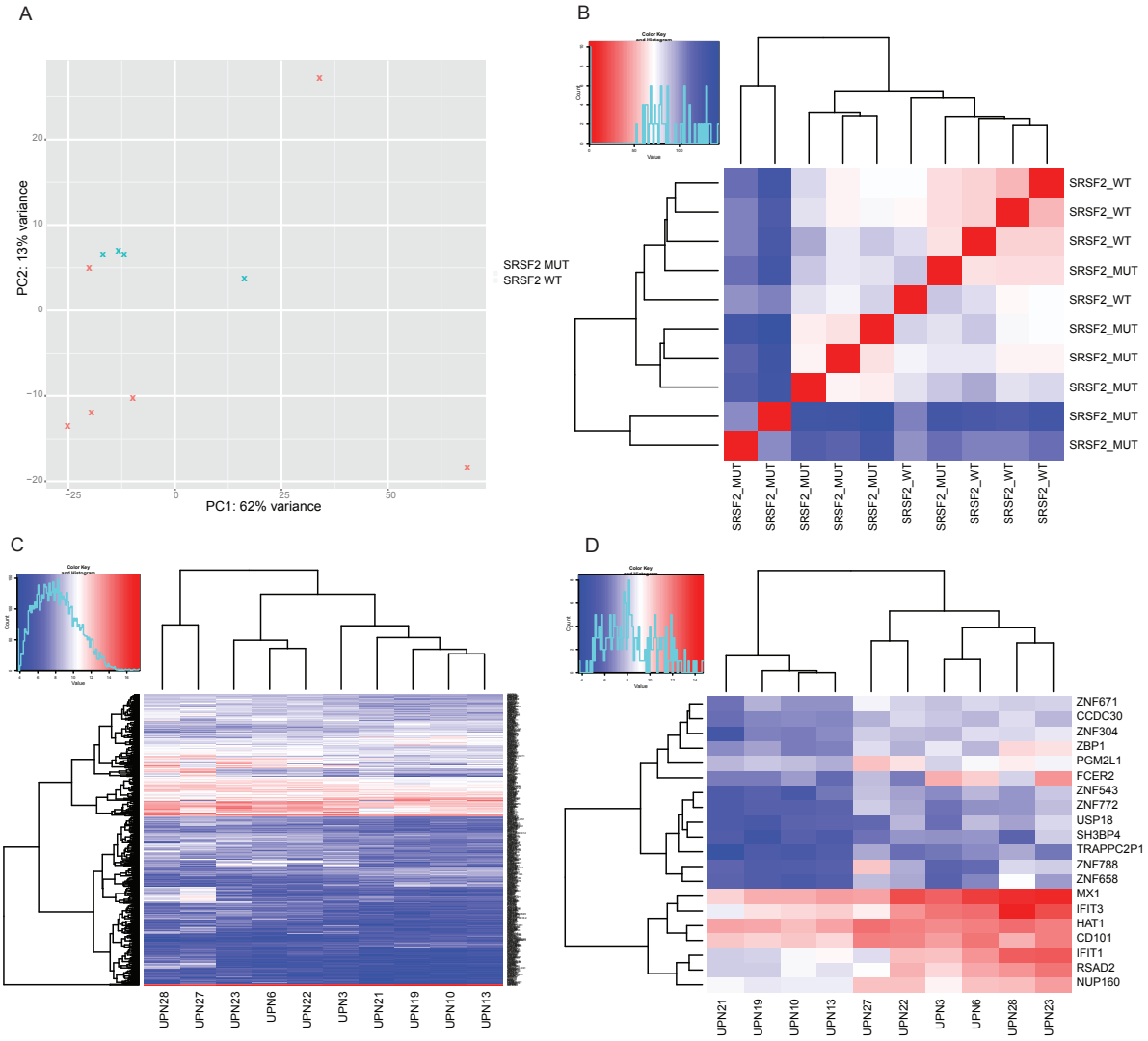


FIGURE 6.6 – *Clusterisation des échantillons de patients et contrôles. A : ACP, B : Matrice de similarité, C : Heatmap des 600 gènes les plus différentiellement exprimés, D : Heatmap des 20 gènes les plus différentiellement exprimés*

les échantillons ne montre pas non plus de séparation suivant la mutation. De la même façon, une analyse non supervisée réalisée sur les 600 (figure 6.6C) ou 500 (figure A.8A) gènes les plus différemment exprimés ne montre pas de regroupement suivant le statut de *SRSF2*. En revanche, une analyse non supervisée réalisée sur 100 (figure A.8B), 400 ou quelques dizaines de gènes (ici 20) les plus différemment exprimés (6.6D) permet une disjonction entre patients mutés et patients non mutés. Le niveau d'expression de quelques gènes semble varier suivant le statut de *SRSF2*. Ceci reste à valider sur des cohortes plus importantes.

Nous avons étendu l'analyse des événements d'épissage à la comparaison des patients mutés et non mutés. Le tableau 6.3 indique le nombre d'événements détectés pour quatre comparaisons. Le nombre de gènes altérés pour chaque comparaison est :

— 674 pour patients *versus* contrôles
— 523 pour patients mutés *versus* contrôles
— 1245 pour patients WT *versus* contrôles

Epissages alternatifs	ExonSkip.	Accept.	Donn.	MultiE.Skip.	MutualExcl.
(Mut+WT) vs Ctrl	593	4	6	194	8
Mut vs WT	27	6	3	14	1
WT vs Ctrl	1022	107	101	338	9
Mut vs Ctrl	450	5	11	135	3

TABLE 6.3 – *Nombre d'événements d'épissage en fonction des comparaisons étudiées*

— 51 pour patients mutés *versus* patients WT.

La comparaison directe des patients suivant le statut de *SRSF2* met en évidence une cinquantaine d'événements, alors que la comparaison indirecte semble en indiquer davantage. Afin de caractériser au mieux les événements de ces comparaisons, nous avons comparé (figure 6.7) d'une part les gènes affectés dans les comparaisons :
— patients *versus* contrôles
— patients mutés *versus* contrôles
— patients WT *versus* contrôles
et d'autre part :
— patients mutés *versus* patients WT
— patients mutés *versus* contrôles
— patients WT *versus* contrôles.

La majorité des événements est retrouvée dans les comparaisons des trois groupes de patients *versus* les sujets contrôles, ce qui indique qu'indépendamment de la mutation de *SRSF2*, les patients ont une importante dérégulation de la machinerie d'épissage.

Un mécanisme encore inconnu d'altération de l'épissage pourrait induire un dérèglement stochastique de l'épissage, surtout en l'absence de la mutation de *SRSF2*. On observe :
— 597 événements spécifiques à la comparaison patients WT *versus* contrôles
— 225 événements communs aux comparaisons patients *versus* contrôles et patients WT *versus* contrôles.

La figure 6.8 indique les gènes dont l'épissage semble modulé par la mutation de *SRSF2*, ceux que l'on retrouve soit exclusivement dans la comparaison patients mutés *versus* patients WT, soit dans l'intersection des comparaisons patients mutés *versus* patients WT et patients mutés *versus* contrôles. Nous avons relevé RIT1 en particulier, qui est sur-exprimé chez les patients (figure 6.4).

6.5 Comparaison des fréquences alléliques des mutations dans l'adn et l'arn des patients

Nous avons finalement cherché à identifier les mutations géniques à partir des données de séquençage d'*ARN*. Les mutations géniques de dix-sept patients (les 10 patients étudiés par ribodéplétion et 7 des 9 patients étudiés par sélection polyA pour l'impact du traitement, les 2 autres patients étant communs aux deux cohortes) ont ainsi été analysées dans leur *ADN* et leur *ARN*. Les résultats ont montré que 20 à 50% des mutations géniques identifiées par séquençage de l'*ADN* ont lieu dans des gènes qui ne sont pas exprimés. Pour une estimation relativement précise des fréquences alléliques, nous avons représenté uniquement les gènes dont les mutations étaient couvertes par au moins 30 lectures dans les échantillons tumoraux d'*ADN* et d'*ARN* (figure 6.9). Seuls quelques gènes présentent une différence de fréquence allélique dans l'*ADN* et l'*ARN*, la majorité des mutations sont retrouvées dans les mêmes proportions. Cette étude doit être poursuivie sur des échantillons séquencés à une grande profondeur.

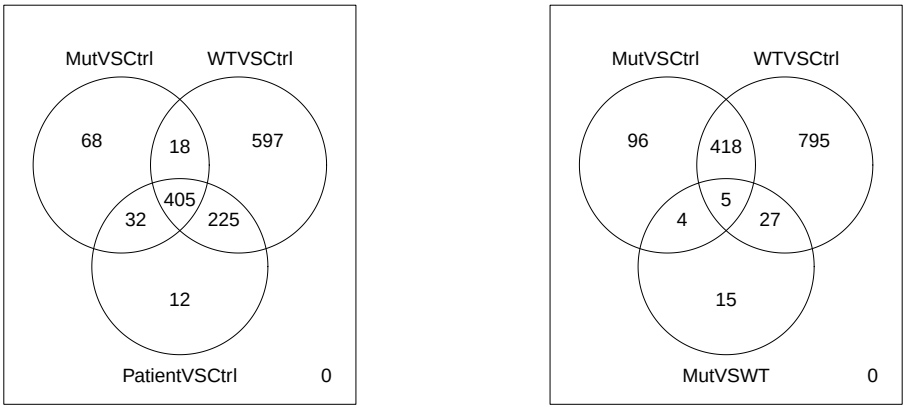


FIGURE 6.7 – *Comparaison des événements d'épissage retrouvés dans les comparaisons : patients mutés versus contrôles, patients WT versus contrôles, patients versus contrôles et patients mutés versus patients WT*

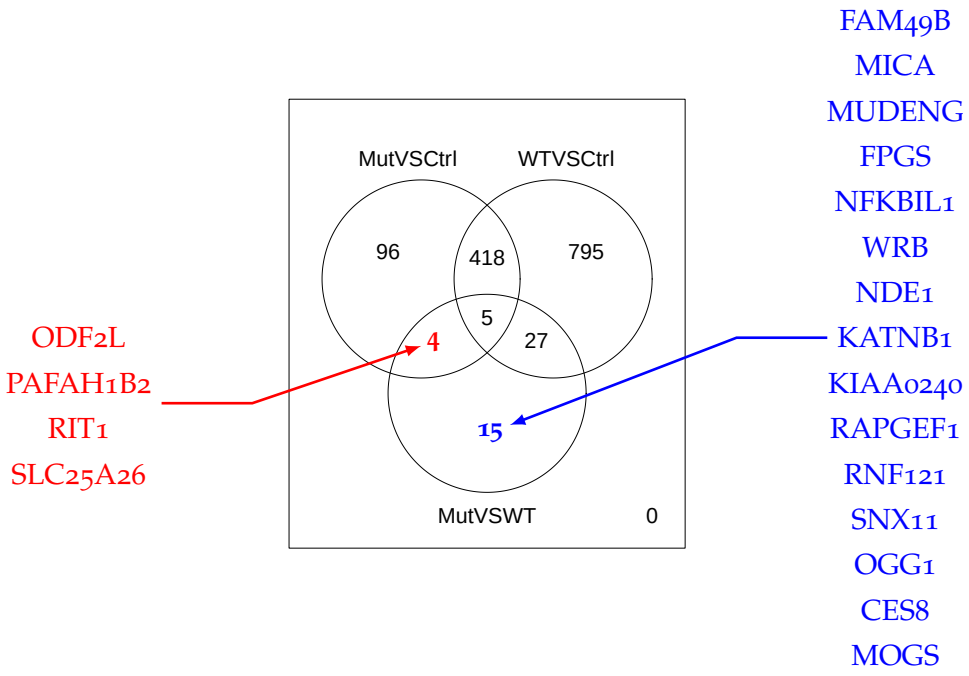


FIGURE 6.8 – *Gènes pouvant être épissés différemment suite à la mutation de SRSF2*

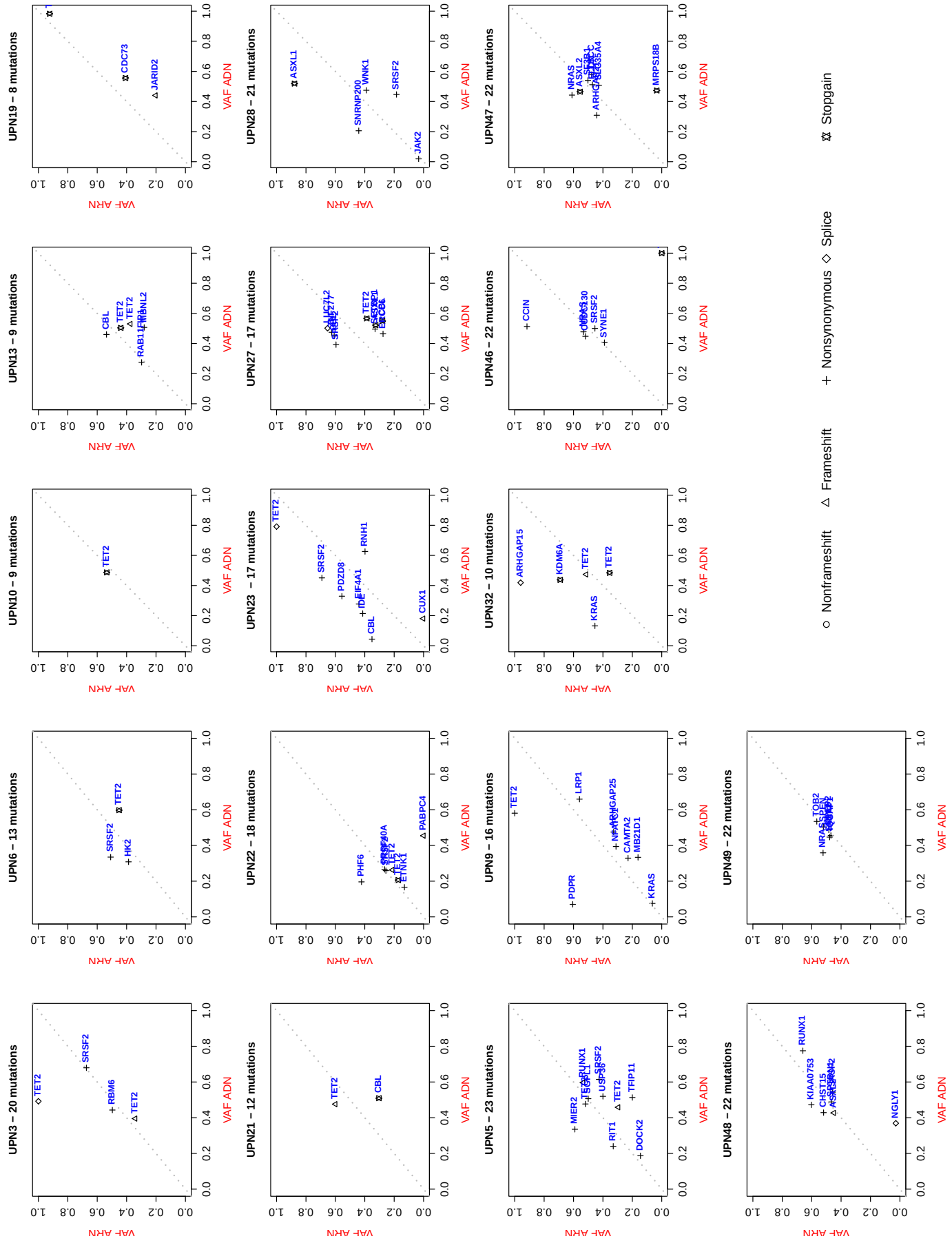


FIGURE 6.9 – *Fréquence de l'allèle variant dans l'ADN et l'ARN des mutations détectées en WES couvertes par au moins 30x*

Quatrième partie

Conclusions et Perspectives

Conclusions et perspectives

7

Le séquençage de l'*ADN* de 65 patients atteints de leucémie myélomonocytaire chronique a permis de conclure que les monocytes de ces patients contiennent en moyenne 14 mutations somatiques dans les régions codantes et 475 mutations somatiques dans les régions non répétées du génome des patients. Ces valeurs situent la leucémie myélomonocytaire chronique au même niveau que les leucémies aiguës myéloïdes ou les myélofibroses, au-dessus de la leucémie myélomonocytaire juvénile (une mutation isolée est souvent suffisante) et bien en deçà de la plupart des tumeurs solides (Alexandrov et al. (2013a), Weinhold et al. (2014)). Parmi les 14 mutations somatiques, environ trois quarts ne sont pas récurrentes et entre 20 et 50% affectent des gènes non exprimés dans les monocytes comme dans les populations cellulaires plus immatures. En moyenne, environ 3 mutations par patient affectent des gènes mutés de façon récurrente, considérés comme potentiellement drivers. Il semble que ce nombre se situe entre 2 et 6 dans la plupart des cancers (Kandoth et al. (2013)).

Notre analyse ne révèle pas de nouveau gène muté de façon récurrente à une fréquence de plus de 10%, mais elle clarifie le paysage des altérations génétiques dans la leucémie myélomonocytaire chronique. Il est désormais improbable qu'il existe des mutations récurrentes très fréquentes qui n'aient pas été identifiées, puisque l'analyse a été faite chez 49 patients, à une couverture moyenne de 112x. Nous confirmons aussi qu'il ne semble exister aucune mutation spécifique de cette maladie.

Dans plus de 90% des leucémies myélomonocytaires chroniques, il existe au moins une mutation affectant un gène régulateur épigénétique, dans 75% un gène codant un facteur d'épissage et dans 60% un gène affectant la signalisation intracellulaire. Nos résultats antérieurs suggèrent que les mutations s'accumulent très souvent, mais pas exclusivement, dans cet ordre (Itzykson et al. (2013b)). L'ordre d'apparition des mutations est vraisemblablement un facteur de variabilité dans l'expression clinique et biologique de la maladie, comme cela vient d'être montré à propos de *JAK2* et *TET2* (Ortmann et al. (2015)).

Notre analyse révèle de nouvelles mutations récurrentes de faible incidence dans 8 gènes exprimés dans les monocytes ou dans les cellules plus immatures CD34+. Les parties codantes de ces huit gènes ont été séquencées dans une cohorte additionnelle de 180 patients *LMMC* et nous avons trouvé les fréquences de mutations suivantes : *PHF6* (7.3%), *NF1* (6.1%), *ETNK1* (3.3%), *DOCK2* (2.1%), *ABCC9* (2.1%), *LUC7L2* (1.7%) et *HUWE1* (1.3%). Des mutations somatiques d'*ASXL2* n'ont pas été retrouvées dans la cohorte additionelle. Pendant ma thèse, des mutations récurrentes d'*ASXL2* ont été identifiées dans les leucémies aiguës myéloïdes (Micol et al. (2014)) et des mutations des gènes *ETNK1* (Gambacorti-Passerini et al. (2015)) et *LUC7L2* (Singh et al. (2013)) ont été rapportées dans la leucémie myélomonocytaire chronique. Séquencer plus d'exomes de patients permettrait très certainement de déceler des mutations rares, comme *EZH2*, qui pourraient être ciblées par des inhibiteurs spécifiques.

Les gènes fréquemment mutés dans la leucémie myélomonocytaire chronique ont aussi été décrits dans les néoplasmes myéloprolifératifs, syndromes myélodysplasiques, leucémies aiguës myéloïdes. Le nombre important de patients séquencés pour tout type de cancers a permis de rechercher la fréquence de mutations des gènes liés aux leucémies et lymphomes, en considérant l'échantillon de sang habituellement contrôle comme tumoral et l'échantillon habituellement tumoral comme contrôle. Comme indiqué dans l'introduction du chapitre 6, plusieurs études récentes convergent pour montrer que des clones de cellules hématopoïétiques mutées pour *TET2* (Busque et al. (2012)), *ASXL1*, *TP53*, *SF3B1*, *ASXL2*, *JAK2* et *SH2B3* peuvent être identifiés dans le sang de sujets sans hémopathie (Xie et al. (2014)). Leur nombre augmente avec l'âge (Xie et al. (2014), Jaiswal et al. (2014), Genovese et al. (2014)). On parle maintenant de "clonalité myéloide de signification indeterminée" (Steensma et al. (2015)). La fréquence de ces mutations augmente avec l'âge : les mutations de *JAK2* semblent apparaître plus souvent chez les quadragénaires (en moyenne 1.2% des quadragénaires auraient des mutations clonales), celles de *DNMT3A* chez les sexagénaires (2.2% de mutations clonales), celles d'*ASXL1* et *TET2* chez les octogénaires (6.1% de mutations clonales). Il est de plus en plus vraisemblable que les mutations dans certains gènes (*DNMT3A*, *TET2*, *ASXL1*, peut-être *SF3B1*) induisent une hématopoïèse clonale sur laquelle la survenue d'autres mutations va générer un phénotype d'hémopathie, myéloïde s'il s'agit de *JAK2* ou de *RAS*, lymphoïde s'il s'agit de *MYD88* ou de *RAF*, aiguë myéloïde s'il s'agit d'*IDH1* ou de *NPM1*. Nous ignorons encore quelle proportion des individus "sains" portant des mutations clonales vont développer une pathologie hématologique à plus ou moins long terme.

La signification des nombreuses altérations somatiques détectées dans les régions non codantes du génome reste inconnue. Les mutations somatiques que nous avons détectées sont essentiellement intergéniques et introniques et sont majoritairement des *SNV*. La signature mutationnelle de ces variants a permis d'identifier trois processus à l'œuvre dans la leucémie myélomonocytaire chronique. Deux sont observés chez les 17 patients et sont retrouvés dans beaucoup d'autres cancers. Il s'agit des signatures 1 et 5 (Alexandrov et al. (2013a)), observées également dans les cellules non tumorales, qui caractérisent le vieillissement. Deux des 17 patients arborent une troisième signature, jamais encore identifiée par l'équipe de Ludmil Alexandrov. Le processus à l'origine de ces mutations n'est pas connu. Il est caractérisé par des mutations C :G>T :A en CpCpC et CpCpT et par un biais transcriptionnel important. Dans les deux cas, les patients ont développé simultanément deux hémopathies distinctes. Le ratio $\frac{Ti}{Tv}$ est de 1.96, ce qui est très proche de celui des génomes non tumoraux, et rejoint les signatures évocatrices du vieillissement. L'analyse de grandes séries de patients sera nécessaire pour aller plus loin. Quoiqu'il en soit, ces données confortent l'impression selon laquelle la leucémie myélomonocytaire chronique est la conséquence de l'accumulation de mutations géniques dans un contexte de vieillissement cellulaire. Ceci est renforcé par le fait que dans les parties codantes du génome, le ratio $\frac{Ti}{Tv}$ est de 2.7, ce qui est très proche de celui des exomes non tumoraux.

L'étude des régions hotspots dans le génome n'a pas permis de déceler de mutations récurrentes à une fréquence élevée dans des régions clés : promoteurs, enhancers, introns. Les anomalies récurrentes dans des régions non codantes peuvent avoir une importance capitale dans le processus de transformation ou de progression tumorale. C'est le cas des mutations récurrentes du promoteur de *TERT* identifiées dans les mélanomes puis dans d'autres tumeurs (Vinagre et al. (2013)) dont les hépatocarcinomes (Pilati et al. (2014)) et des mutations des enhancers, comme celui de *PAX5* (Puente et al. (2015)). Cependant, seuls 17 patients ont été analysés à une couverture moyenne de 30x et 45% du génome a été éliminé car étant constitué de régions répétées. Dans ces conditions, il est possible qu'il existe des mutations somatiques récurrentes qui soient passées inaperçues. Il serait donc utile de séquencer davantage d'échantillons à une profondeur plus importante en éliminant moins de régions répétées (10% serait possible) lors de l'analyse. Une telle analyse permettrait de valider les processus mutationnels identifiés et potentiellement d'en

discerner de nouveaux, au-delà peut-être de l'hypothèse du seul vieillissement.

Nous n'avons pas pu réaliser la recherche de variations somatiques structurales de grandes tailles dans les données de génome. Des événements au potentiel driver pourraient exister, comme des protéines de fusion, même si les analyses préliminaires réalisées sur les 5 patients pour lesquels nous disposons de fibroblastes cutanés comme contrôles n'en ont pas montré. Nous pourrions analyser plus précisément les sites fragiles répertoriés dans les différents types cellulaires. Les patients *LMMC* pourraient présenter des anomalies au niveau de ces sites sensibles, comme le locus 16q23.3-24.1 identifié dans les cancers du sein, de l'ovaire et de la prostate (Bednarek et al. (2000)).

L'analyse des régions répétées est complexe du fait des problèmes d'alignement des lectures relativement courtes. Ces régions comportent notamment les transposons/rétrotransposons, zones dans lesquelles peuvent se trouver des anomalies. Il existe plusieurs exemples de maladies génétiques dont l'origine se situe dans des zones répétées. La maladie de Huntington par exemple, consiste en une répétition de 3 nucléotides dans le gène *IT15* situé sur le bras court du chromosome 4. Similairement, la dystrophie myotonique de Steinert est une maladie héréditaire due à une expansion instable (au moins 37 répétitions) d'un triplet du gène *DMPK*. Le nombre de répétitions du triplet est souvent associé à la sévérité de la maladie ainsi qu'à l'âge d'apparition des symptômes. Contrairement au cas de la leucémie myélomonocytaire chronique, ces anomalies sont héréditaires et un phénomène d'anticipation peut se produire d'une génération à l'autre. Plusieurs cancers sont dus à des intégrations de virus dans ces régions répétées, comme le virus d'Epstein-Barr ou le papillomavirus.

Les cancers contiennent de quelques centaines à quelques dizaines de milliers de mutations somatiques. Il est couramment admis que seules quelques-unes sont à l'origine de cancers, alors que la majorité des événements ne contribue pas au phénotype. Les drivers confèrent des phénotypes avantageux aux cellules tumorales, augmente leur compétitivité par rapport aux cellules non mutées, au moins dans un contexte donné. Cette propriété vient de l'effet sur les voies liées au cancer, l'occurrence élevée des mutations dans des voies communes, gènes ou hotspots. Les drivers augmentent la population tumorale en augmentant le taux de divisions cellulaires (par une mutation activatrice de *KRAS* par exemple) ou en diminuant la mort cellulaire (par knockout de *TP53* par exemple). Un driver conduit à l'expansion clonale de la sous-population portant ce driver, conduisant à une rapide croissance cellulaire pendant une courte période.

McFarland et al. (2013) ont étudié l'impact des mutations passagères. Celles-ci représentent la majorité des mutations somatiques observées. Ces mutations sont considérées comme n'ayant pas de rôle dans le cancer (neutres), car elles ne sont pas récurrentes et sont dispersées à travers le génome. Cependant, de nombreuses mutations passagères se trouvent dans des gènes codants des protéines et d'autres éléments fonctionnels et peuvent avoir des effets délétères pour les cellules tumorales. Bien que les mutations passagères délétères soient évincées par sélection négative, les mutations passagères à effet modéré peuvent échapper à la sélection négative et s'accumuler. Même si individuellement leur effet est faible, leur effet cumulé altère la progression, conduisant à plusieurs phénomènes oncologiques difficiles à expliquer avec l'approche traditionnelle basée sur les drivers. La croissance de la tumeur s'arrête quand l'effet d'un ou de plusieurs drivers est contre-balancé par le taux de mort cellulaire. Pendant que la population attend la survenue d'un driver additionnel, les mutations passagères s'accumulent, induisant une diminution graduelle de la population tumorale. La probabilité soit d'une croissance incontrôlée, soit d'une régression spontanée dépend de la taille de la population tumorale : les tumeurs affectant un nombre important de cellules progressent le plus souvent, alors que les plus petites tumeurs ont tendance à régresser. Les grandes populations acquièrent des drivers plus fréquemment, puisqu'il y a plus de

cellules dans lesquelles ces drivers peuvent apparaître. Les mutations passagères à effet modéré, bien plus que celles à effet délétère ou neutre, ont un effet majeur sur la progression tumorale. En provoquant son ralentissement, elles peuvent s'accumuler en grand nombre.

Les mutations drivers sont distinguées des autres mutations par comparaison de la fréquence de mutations conduisant à une modification de protéines avec la fréquence de mutations synonymes dans des mêmes gènes. Supek et al. (2014) ont étudié le rôle, dans les cancers, de ces mutations synonymes, bien souvent mises de côté. Ils ont analysé plus de 3000 exomes et plus de 300 génomes tumoraux. Les positions synonymes dans une séquence génique permettent le stockage d'informations telles que la vitesse ou la précision à ou avec laquelle un *ARNm* est traduit, comment un *ARNm* est exprimé ou épissé, ou comment une protéine est exprimée. Ces mécanismes, entre autres, démontrent que les changements dans les séquences géniques qui sont synonymes pour la séquence protéique, ne sont pas sans effet sur la fonction.

Les oncogènes peuvent être activés par mutation nonsynonynome, amplification ou translocation, tandis que les gènes suppresseurs de tumeur sont inactivés par délétion, perte d'hérérozygotie, gain de codon stop, *INDEL* décalant le cadre de lecture ou mutation dans un intron inactivant les sites d'épissage d'ARN pré-messager. Les oncogènes présentent un nombre de mutations somatiques synonymes significativement élevé dans les génomes tumoraux, contrairement aux gènes suppresseurs de tumeur (Supek et al. (2014)). Les mutations synonymes sont rares dans ces gènes, à l'exception peut-être de *TP53*. Les mutations synonymes dans les oncogènes sont plus fréquentes dans les cancers présentant le moins de mutations (leucémies, cancers du sein et de l'ovaire). Elles ciblent préférentiellement les régions conservées ou certaines régions spécifiques, ainsi que les motifs d'épissage exonique avec très souvent un épissage anormal : la moitié des drivers synonymes putatifs est associée à des anomalies d'épissage. La fréquence de mutations nonsynonymes dans certains oncogènes varie selon les tissus. Par exemple, l'oncogène *BRAF* est fréquemment muté dans les mélanomes alors que *KRAS* est fréquemment muté dans les cancers colorectaux. Environ 1 mutation synonyme sur 5 dans les oncogènes connus à ce jour a été sélectionnée, et cette proportion atteint environ 1 mutation synonyme sur 2 dans les oncogènes caractéristiques d'un type de cancer (Supek et al. (2014)). Tous ces éléments démontrent l'impact d'une partie des mutations synonymes et incitent les études de génomes tumoraux à venir à rechercher ces enrichissements en mutations synonymes.

L'évolution des mutations détectées dans la première partie du travail a été étudiée pour 17 des 49 patients analysés en WES à plusieurs temps. Les mutations acquises dans les parties codantes du génome ne disparaissent pas, à l'exception de quelques sous-clones, et continuent à s'accumuler lentement chez les patients, qu'ils soient traités ou non et qu'ils répondent ou non à leur traitement. Les patients répondant à leur traitement ne voient pas disparaître leurs mutations et peuvent même continuer à en accumuler.

L'étude menée par Craddock et al. (2013) suggère que le traitement d'une leucémie aiguë myéloïde par Azacytidine ou acide valproïque ne réduit le nombre de cellules souches ou progénitrices leucémiques que chez les sujets répondeurs et ne permet pas l'éradication complète de ces cellules. Ces populations se ré-expandent au moment des rechutes. De la même manière que nous n'observons pas l'éradication des mutations, les populations leucémiques ne sont pas éliminées.

Afin de déterminer si l'effet limité des traitements actuels est dû à la présence de mutations et si l'évolution de la maladie est corrélée à l'évolution des mutations, il faudra poursuivre notre analyse sur plus de patients et sur une longue durée. Ces études permettront de préciser le

taux de nouvelle mutation par an et de préciser si ce taux varie en fonction du traitement et du statut (répondeur, stable, non répondeur). D'autres variables pourraient être étudiées telles que l'influence du contexte clinique et biologique : est-ce que ce taux de mutation est le même suivant le type de leucémie myélomonocytaire chronique (*LMMC*1, *LMMC*2, *LMMC* induite) ? Cette analyse pourrait être conduite au niveau du génome entier : observe-t-on le même comportement à l'échelle maximale que l'on puisse analyser ? Est-ce que le temps modifie la signature mutationnelle des patients à l'échelle de l'exome et du génome ?

La corrélation entre mutations et maladie pourrait être complétée par l'étude des tissus non cancéreux de patients *LMMC*. Ces tissus non cancéreux peuvent acquérir des mutations au cours de la vie de l'individu. Nous avons détecté 14 mutations somatiques et 3 gènes *a priori* drivers mutés en moyenne dans les monocytes des patients. Les deux patients pour lesquels nous avons analysé également des $CD34^+$ portent les mêmes mutations dans ces cellules plus immatures. Notre analyse séquentielle dans les monocytes sur un nombre de cas limité n'a pas montré une forte accélération de l'acquisition des mutations chez les patients. De ce fait, si on élimine les 3 gènes mutés *a priori* drivers qui sont spécifiques du clone leucémique, il doit ne rester que les mutations accumulées au cours de la vie de l'individu. Ainsi, y a-t-il le même nombre de mutations somatiques dans des tissus non tumoraux des patients que dans les cellules tumorales dont on soustrait les mutations drivers, soit environ 11 mutations ? Pour répondre, on pourrait refaire les analyses en inversant les échantillons contrôles et tumoraux, avec un échantillon contrôle présentant un taux de division cellulaire similaire. Cette question peut-être étendue à l'ensemble du génome.

Nous avons étudié l'effet du traitement sur une cohorte de 9 patients, à deux temps, sur les aspects épigénétiques, par séquençage d'*ARN* polyadénylés et par séquençage d'*ADN* bisulfite (*ERRBS*). Au premier temps, les 9 patients étaient non traités, au deuxième temps, 3 patients restaient non traités et les 6 autres étaient traités par agents déméthylants (Décitabine ou Azacitidine). Trois des patients étaient répondeurs et les trois autres étaient des patients stables, *i.e.* sans amélioration ni dégradation de l'état clinique. Nous avons d'abord étudié l'expression génique chez nos patients. Chez les patients répondeurs, nous avons identifié environ 500 gènes dont l'expression était dérégulée entre les 2 temps, contre une soixantaine seulement chez les patients stables et 0 chez les patients non traités. Nous avons ensuite étudié le niveau de méthylation chez ces mêmes patients. Chez les patients répondeurs, nous avons identifié environ 35000 régions méthylées de manière différentielle entre les 2 temps, contre une centaine seulement chez les patients stables et 1 chez les patients non traités. Finalement, les traitements actuels ont un effet fortement épigénétique chez les répondeurs mais n'influent pas sur les mutations, ce qui pourrait expliquer les rechutes systématiques. Il serait intéressant de mesurer ces effets lorsque les patients ne répondent plus. Est-ce que comme dans l'étude de Craddock et al. (2013) nous observerons le retour aux niveaux initiaux ?

Récemment, Shen et al. (2015) ont recherché les anomalies des cellules $CD34^+$, des granulocytes et des monocytes par immunophénotypage en flux cytométrique avant et après traitement par agents hypométhylants. Les patients présentant une amélioration hématologique ont montré une diminution significative du nombre de blastes et de monocytes. Alors que les patients avec amélioration hématologique ont une expression normale du CD14, une partie des patients non répondeurs a conservé une expression altérée du CD14. Chez trois patients ayant reçu une greffe, les myéloblastes anormaux $CD34^+$ n'étaient plus détectables après la greffe. Shen et al. (2015) suggèrent que, suite à un traitement par agents hypométhylants, les patients peuvent être en rémission, mais qu'il subsiste une maladie résiduelle, ce qui va dans le sens de nos observations.

Pour plus de robustesse, il serait adéquat d'étudier l'effet du traitement sur l'expression génique, le niveau de méthylation et l'épissage (ce dernier n'ayant pu être réalisé sur la cohorte des 9 patients étudiés ici) sur plus de patients. Il serait intéressant de rajouter des patients non répondeurs, afin de voir si la dégradation de l'état clinique induit des anomalies différentes de celles induites par le traitement, ce à quoi on s'attend.

Afin de s'assurer que ces traitements ne favorisent pas l'apparition de mutations, il faudrait analyser la signature mutationnelle avant et après traitement. Pour que ce soit le cas, il faudrait constater l'apparition de nouvelles mutations, ce qui ne semble pas être le cas au niveau de l'exome.

Nous avons observé de très nombreuses régions différentiellement méthylées entre les patients répondeurs au temps non traité et ces patients une fois traités. Il serait intéressant de comparer ce nombre et ces types d'événements aux dérégulations obtenues en comparant des sujets sains à des patients. Le niveau de méthylation des patients traités et répondeurs est-il similaire à celui de sujets sains ? En d'autres termes, le niveau de méthylation a-t-il été normalisé ou est-il encore à un niveau pathologique ?

L'Azacitidine et la Décitabine n'étant pas métabolisées de la même manière, est-ce que le choix des thérapeutiques influence les modifications d'expression génique et de niveau de méthylation de manière différente ? De plus, si les traitements induisent une signature mutationnelle spécifique, celle-ci est-elle dépendante du choix des thérapeutiques ? Y a-t-il un traitement préférable à l'autre ?

Ces changements d'expression et de méthylation ont été observés au moins six mois après le premier cycle de traitement suite à l'administration d'agents déméthylants. Mais quand ces modifications apparaissent-elles ? Devancent-elles la réponse clinique ? Auquel cas, elles pourraient servir de marqueur de réponse, ce qui serait utile étant donné que la réponse clinique, quand elle se produit, est observée au bout de plusieurs semaines ou mois. Si ces modifications étaient rapides (quelques heures ou quelques jours), cela permettrait de distinguer les répondeurs des non répondeurs et d'ainsi éviter aux patients non répondeurs les effets indésirables des traitements pendant plusieurs mois. Malheureusement, rien d'autre ne pourrait leur être proposé à l'heure actuelle.

Enfin, pourquoi les agents déméthylants apportent-ils un bénéfice dans certaines hémopathies uniquement ? Il serait intéressant de comparer les voies affectées par méthylation, dérégulation d'expression génique, mutation, ... chez des patients présentant un bénéfice et chez des patients ne tirant aucun bénéfice de ces traitements sur de grandes cohortes de patients atteints de diverses hémopathies. Ce bénéfice est tout de même limité si on le compare à celui obtenu avec d'autres traitements, comme l'Imatinib par exemple. Dans leur étude des mutations passagères, McFarland et al. (2013) s'intéressent aux approches thérapeutiques, basées sur les mutations passagères, qui ont pour objectif soit d'augmenter leur taux de mutation, soit d'augmenter leur effet délétère. Dans les deux cas, les cellules tumorales régressent. Cette approche pourrait être efficace dans le cas de la leucémie myélomonocytaire chronique.

Les anomalies d'expression et d'épissage sont présentes en grand nombre dans les monocytes de patients dans la leucémie myélomonocytaire chronique. Nous avons mis en évidence des candidats très intéressants, aussi bien au niveau des dérégulations géniques que des événements d'épissage. Les familles de gènes où plusieurs membres sont dérégulés sont également des pistes pour l'avenir, qu'il convient d'étudier dans une plus grande cohorte et dans des cellules plus immatures. L'étude de l'effet de la mutation de *SRSF2* sur le niveau d'expression n'a pas conduit à la détermination de candidats intéressants fortement dérégulés. Les dérégulations d'épissage, quoiqu'apparemment pas très nombreuses, ont mis en exergue *RIT1*. Sont particulièrement intéressants les quelques gènes à la fois anormalement exprimés et anormalement épissés, qui sont parfois mutés : *U2AF1*, *LUC7L2* et *RIT1*. Cette étude ne permet pas de conclure à une absence d'effet de la mutation de *SRSF2* sur le niveau d'expression et sur l'épissage dans la leucémie myélomonocytaire chronique, puisque la profondeur n'était pas suffisante pour tous les échantillons dans notre étude. D'autant plus que certaines études ont rapporté des effets de cette mutation. Komeno et al. (2015) par exemple ont identifié des événements d'épissage chez la souris. Il convient donc d'analyser ces mêmes effets dans des cellules plus immatures, et à une profondeur de séquençage importante pour les événements d'épissage.

Une question commune à toutes les analyses réalisées et à venir est de déterminer la cellule à étudier la plus pertinente pour les patients, celle qui permettra de détecter les anomalies à cibler pour améliorer la survie des patients. Faut-il étudier les cellules matures, susceptibles de contenir la majorité des anomalies, ou au contraire, les cellules immatures, comportant les toutes premières anomalies ? Tout dépend de la question d'intérêt, mais aussi des techniques employées. Le séquençage à très haut débit nécessite un nombre important de cellules (au moins 1 million). Les cellules peu présentes dans l'organisme ne sont donc pas analysables par cette méthode. Quand nous avons commencé le projet en 2011, il nous était impossible de séquencer des $CD34^+$ du sang car ces cellules sont très rares, alors qu'aujourd'hui c'est possible avec la diminution des quantités nécessaires. Les mutations de la leucémie myélomonocytaire chronique s'accumulant dans le temps, rechercher les mutations dans les monocytes nous permet de capter aussi bien les premières que les dernières mutations à être survenues. En revanche, il est possible que les dérégulations géniques observées dans les monocytes soient complètement différentes de celles présentes dans des cellules plus immatures et que les dérégulations pathologiques ne soient visibles qu'à des stades plus immatures.

Il convient finalement de s'interroger sur les analyses réalisées. Était-il et sera-t-il possible de mieux faire ? Plusieurs points peuvent déjà être optimisés. Tout d'abord, il faudrait limiter les régions répétées du génome exclues de l'analyse. Ensuite, les données *WES* peuvent être réalignées avec l'option MEM de l'algorithme BWA, qui est recommandée à l'heure actuelle pour les lectures de plus de 70bp. Les analyses de *LOH* et *CNV* peuvent être refaites en utilisant d'autres outils. A l'heure de la rédaction de cette thèse, l'efficacité de Control-FREEC sur les données d'exomes est remise en question. L'ensemble des données générées dans ce projet devrait être réanalysé tous les deux ou trois ans au fur et à mesure des avancées logicielles afin de vérifier que rien ne nous a échappé à cause de limitations des méthodes d'analyses.

Ce travail a permis d'établir le paysage mutationnel de la leucémie myélomonocytaire chronique. Nous avons pu apprécier la complexité des clones leucémiques des patients, pouvant porter jusqu'à sept drivers dans les régions codantes. Nous avons également pu mettre en évidence de nombreuses dérégulations géniques et des épissages alternatifs, parmi lesquels certains candidats seront étudiés en détails. Enfin, ce projet a surtout permis de compléter les connaissances sur l'effet des agents déméthylants en démontrant une hypométhylation et le manque de cytotoxicité sur les monocytes mutés. Il semble aujourd'hui indispensable et urgent de développer des thérapeutiques agissant sur les mutations.

Ces quatre années de doctorat m'ont permis d'explorer l'analyse de données génomiques de manière globale. J'ai pu analyser divers types de données *NGS* dans le cas de pathologies hématologiques, avec quelques applications en dehors de la leucémie myélomonocytaire chronique. J'ai également pu me familiariser avec le domaine de l'hématologie, ce qui m'a permis d'appréhender les dérégulations essentiellement myéloïdes qui peuvent s'y produire. De manière plus globale, j'ai pu percevoir les problématiques courantes en cancérologie, notamment grâce aux nombreux séminaires organisés à Gustave Roussy.

Annexes

A

A.1 Analyse de séquences d'adn

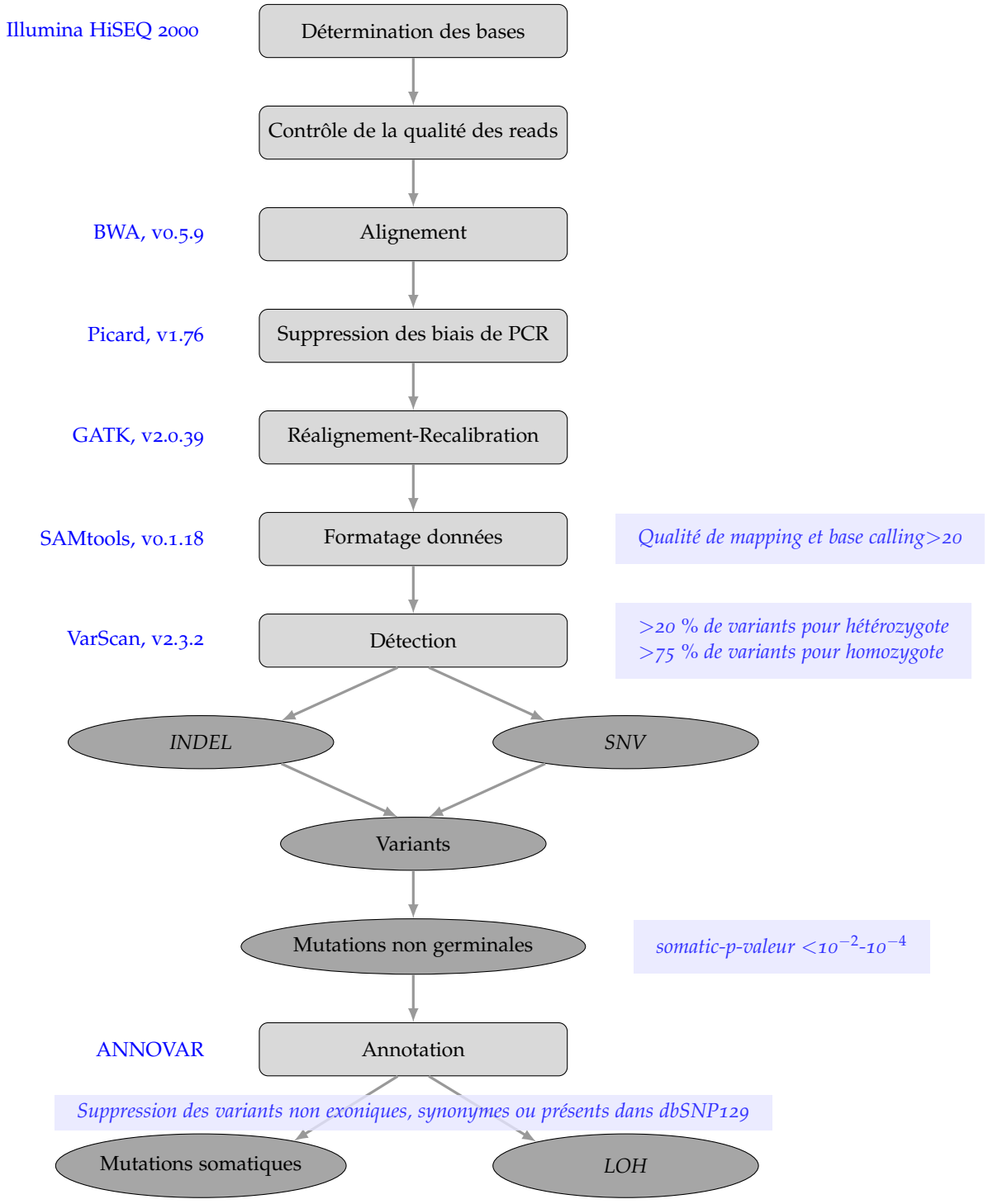


Figure A.1 – *Pipeline utilisé pour l'analyse des données d'exome*

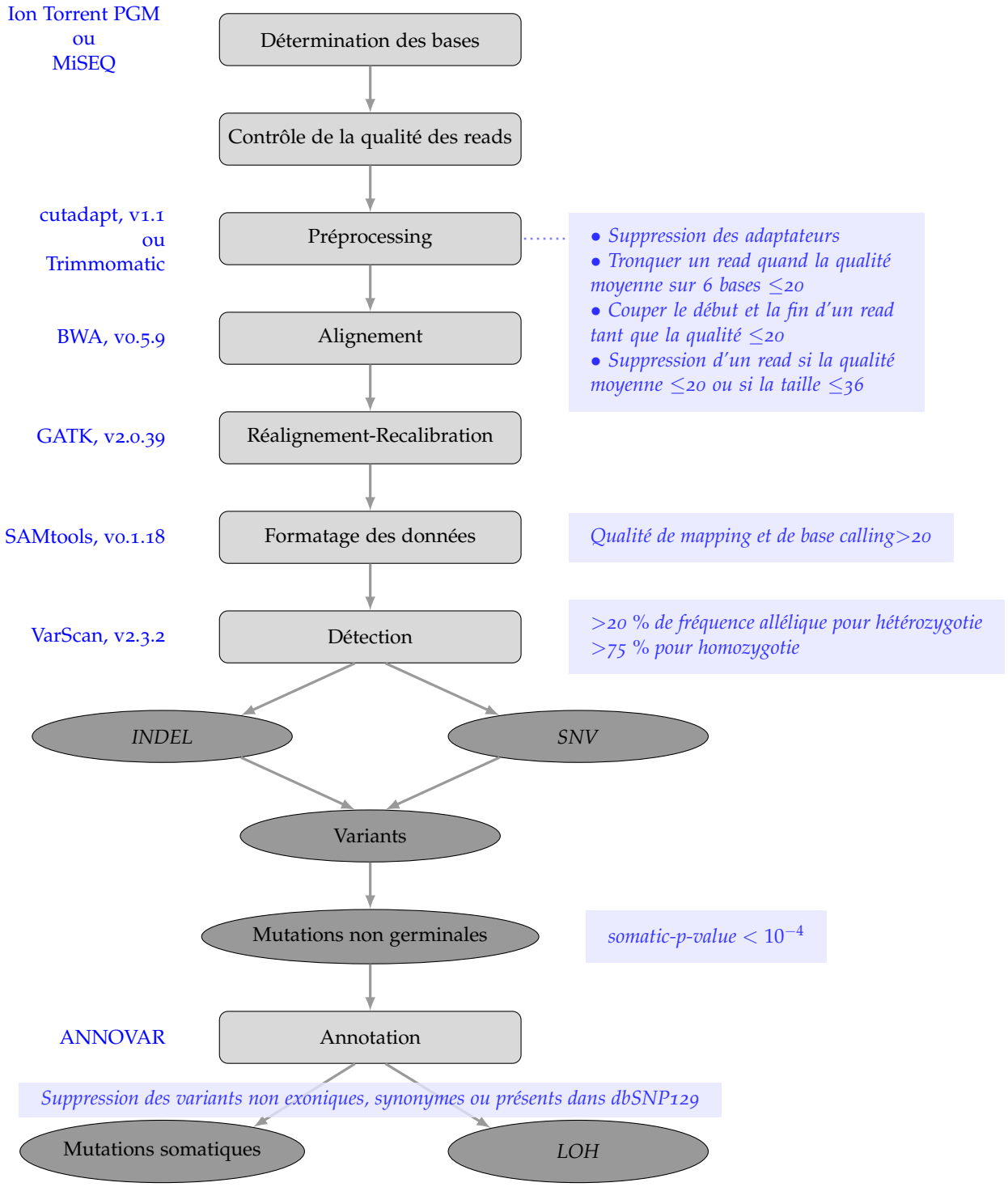


FIGURE A.2 – *Pipeline utilisé pour l'analyse de données de reséquençage (PGM et MiSeq)*

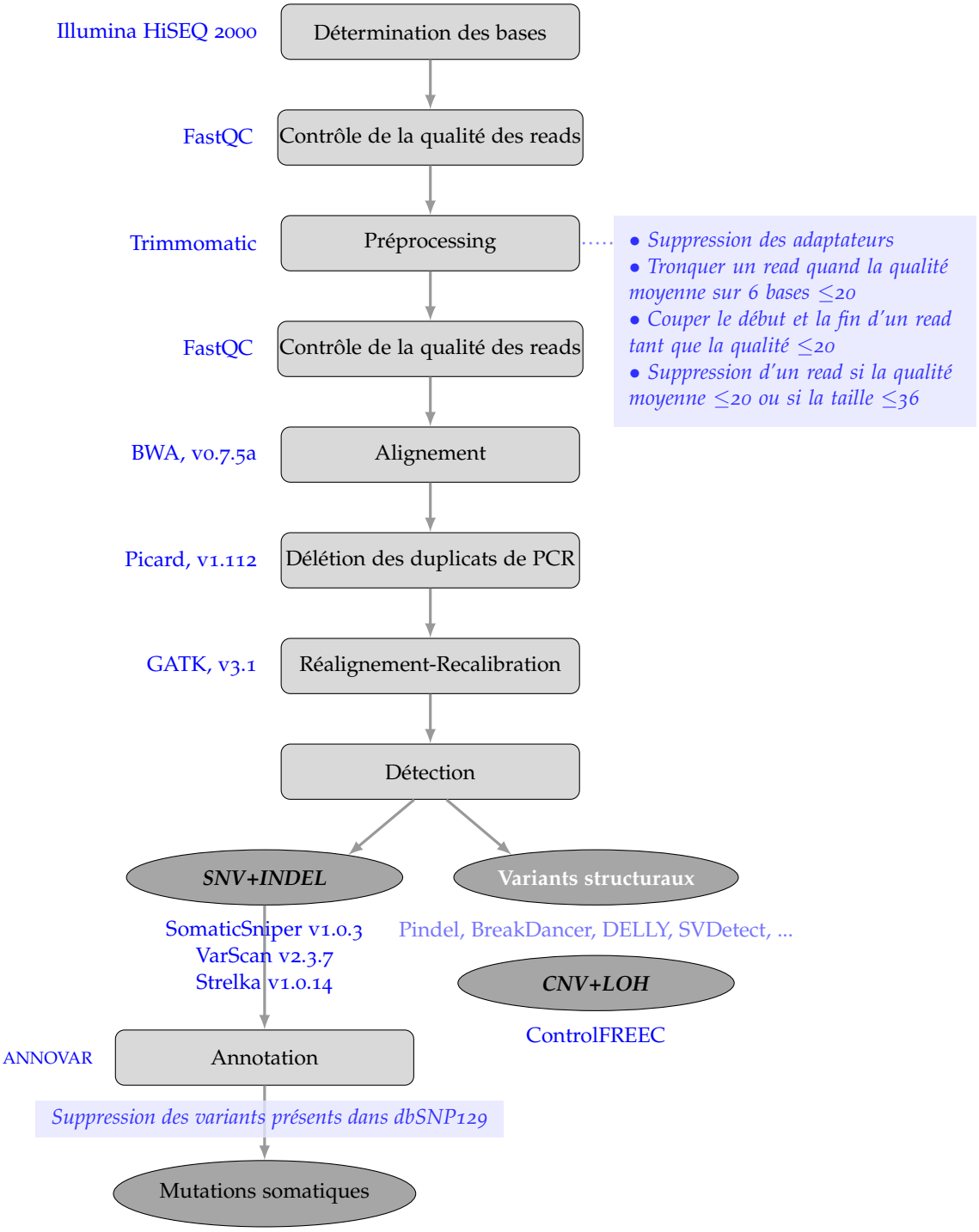
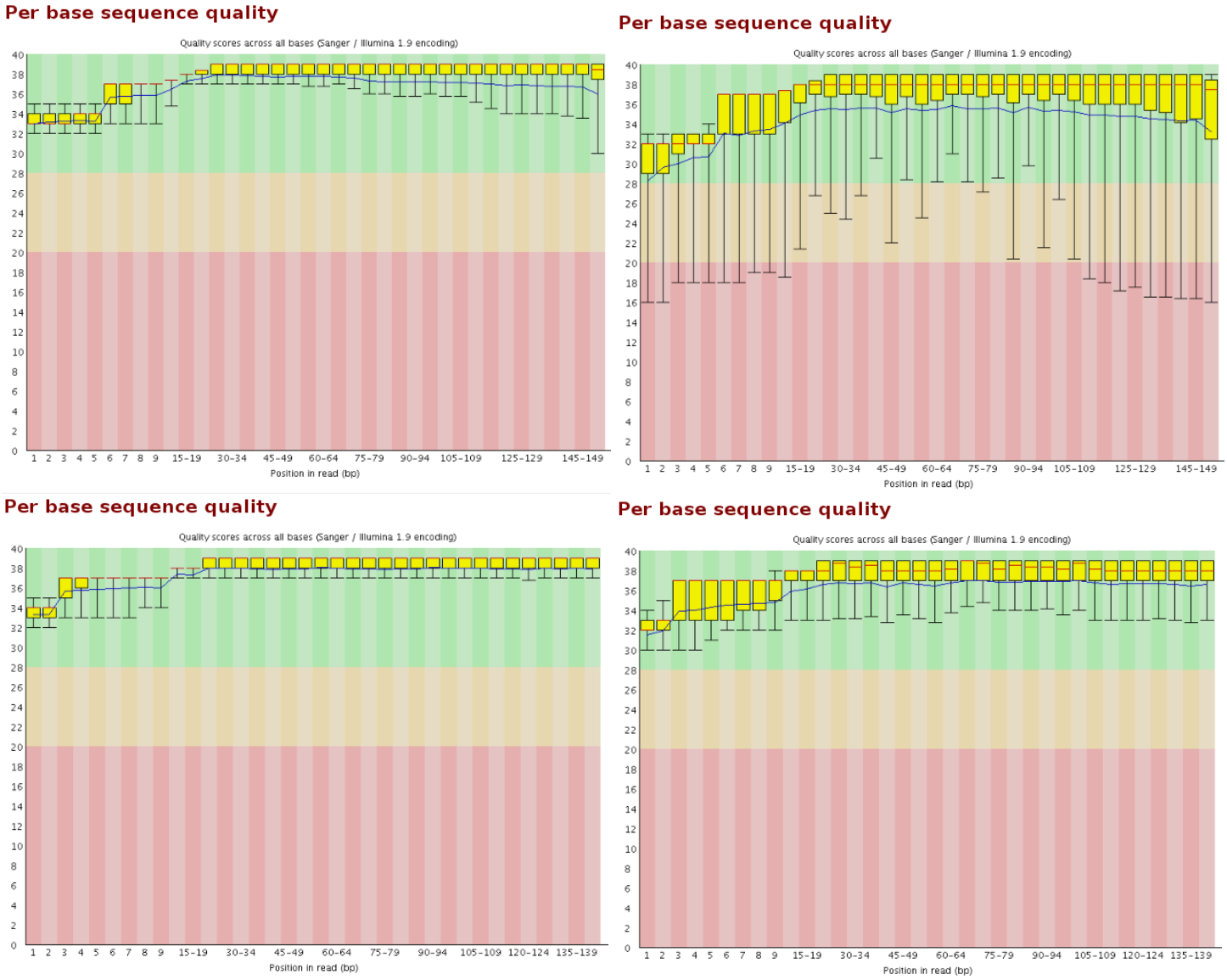


FIGURE A.3 – *Pipeline utilisé pour l'analyse de données de génome*

A.2 Exemples de contrôle qualité et préprocessing des données

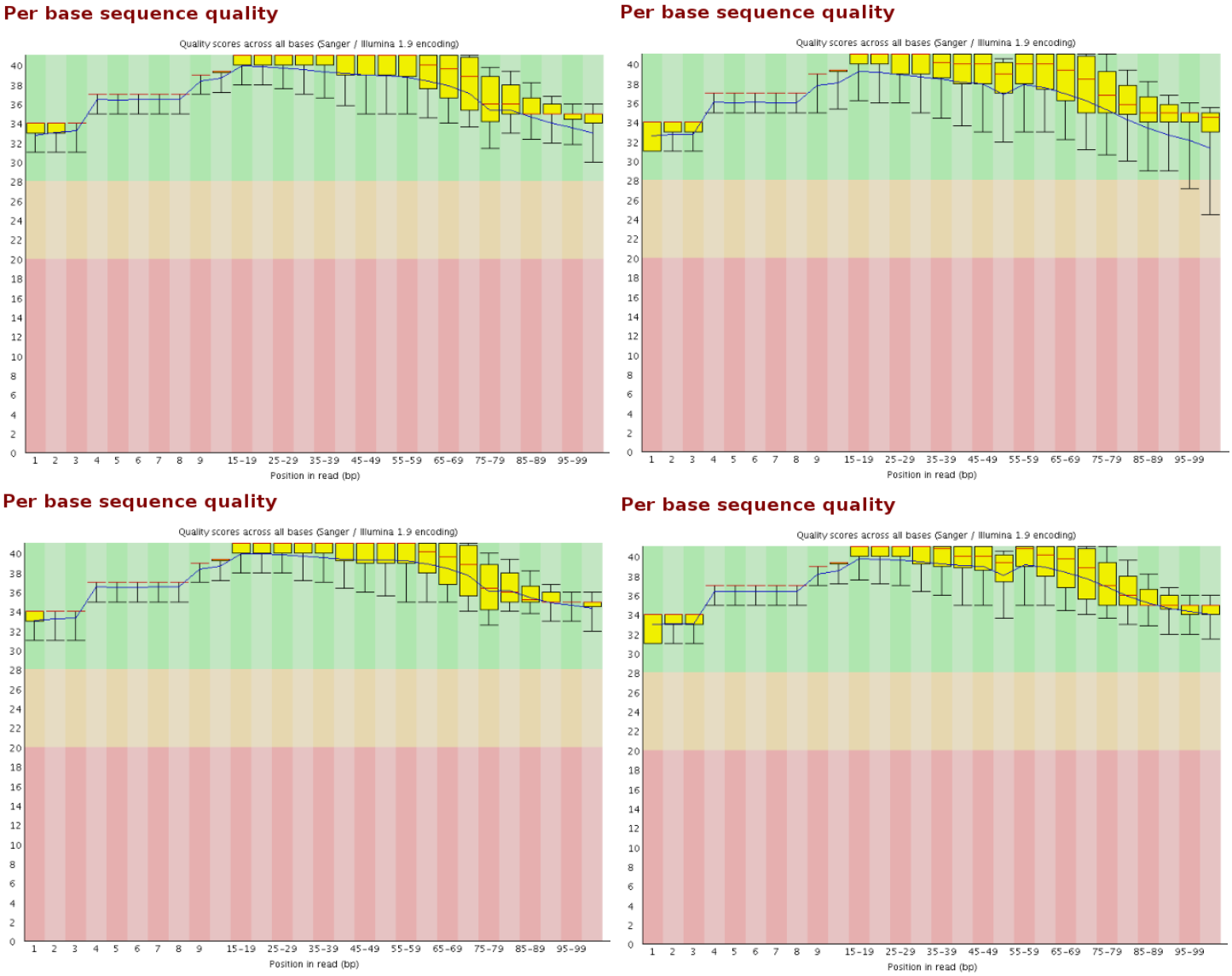
A.2.1 Données de séquençage ciblé



Nous présentons le profil qualité classique d'un séquençage ciblé réalisé avec le séquenceur MiSeq. La taille des reads est de 150 bp. Le profil qualité des données brutes est donné dans la partie supérieure et le profil qualité des données après préprocessing est donné dans la partie inférieure. Le profil des lectures en sens direct est à gauche et en sens indirect à droite. Le même préprocessing est appliqué de manière automatique sur tous les échantillons pour les lectures en sens direct et indirect.

On observe que les séquences en sens direct sont de meilleure qualité que celles en sens indirect. Le préprocessing a mené à supprimé à peine 1% des données en sens direct, alors qu'il en a supprimé environ 30% en sens indirect.

A.2.2 Données de séquençage de génome



Nous présentons maintenant le profil qualité classique d'un séquençage *WGS* obtenu avec Illumina HiSeq2000. La taille des reads est de 101 bp. Le profil qualité des données brutes est donné dans la partie supérieure et le profil qualité des données après préprocessing est donné dans la partie inférieure. Le profil des lectures en sens direct est à gauche et en sens indirect à droite. Le même préprocessing est appliqué de manière automatique sur tous les échantillons pour les lectures en sens direct et indirect.

On observe que les séquences en sens direct sont légèrement de meilleure qualité que celles en sens indirect. Le préprocessing a conduit à supprimer à peine 1% des données en sens direct et environ 3% en sens indirect, bien moins qu'avec le séquençage ciblé.

A.3 Résultats complémentaires sur l'expérience rna-seq avec arn ribodéplétés

A.3.1 Qualité des données

Échantillon	Séquences brutes	Séquences filtrées R1	Séquences filtrées R2	Alignement	Reads sur Transcriptome
UPN3	64764163	64404155	61218764	92.0%	41.23
UPN10	39955124	39499621	37015969	90.8%	25.76
C1	5986853	5888867	4901412	78.3%	3.27
UPN6	140913689	138807351	128695323	86.3%	99.77
UPN22	62486911	61482712	43732663	89.3%	29.71
C2	147272661	145519479	117386488	85.2%	96.54
UPN13	59571444	59094586	56398988	91.1%	36.31
UPN23	69058317	68546792	64893332	91.7%	46.49
UPN27	295200614	291912126	271934130	91.8%	145.81
UPN19	25506075	25259699	23913787	91.2%	16.74
C3	90685564	89506948	81698765	89.1%	57.59
UPN28	64650049	63829332	59523041	88.2%	40.74
UPN21	40334305	39947238	37814838	90.2%	24.99
C4	64191042	63652709	59982832	90.7%	45.29

Table A.1 – *Couverture des échantillons RNASeq ribodéplétés*

A.3.2 Analyse des données

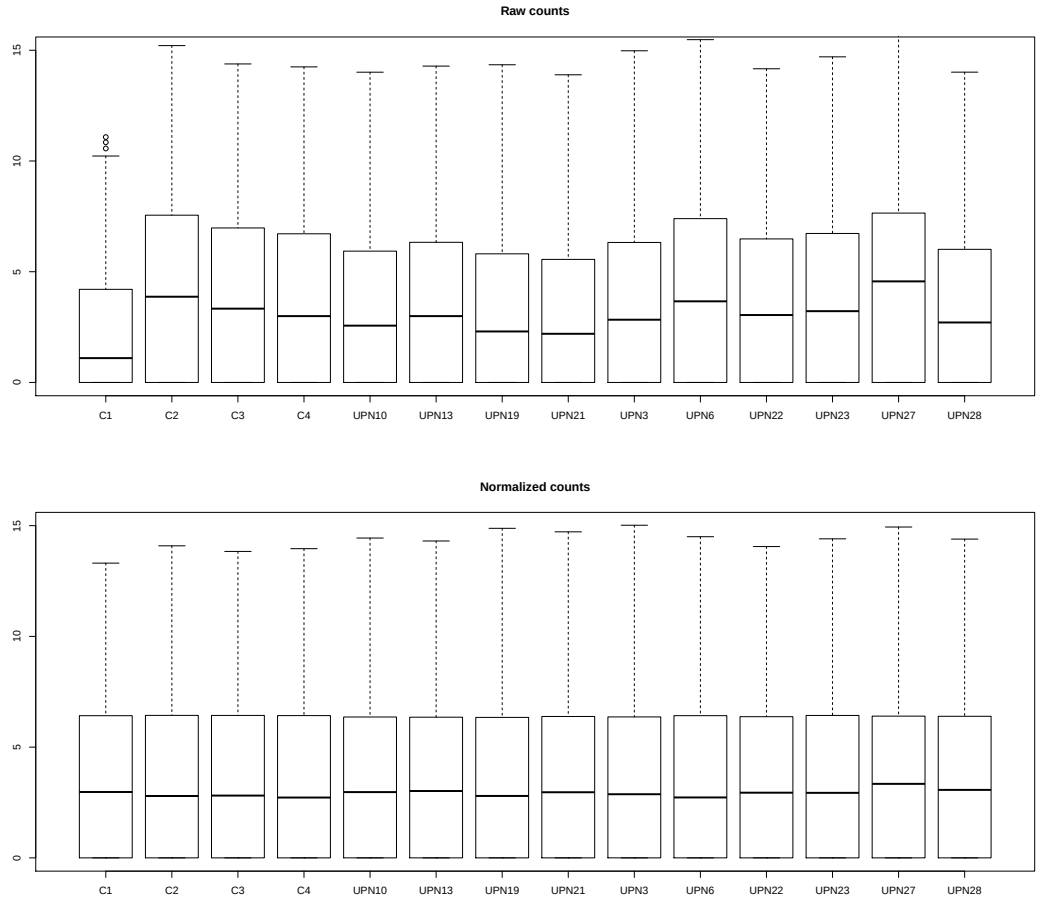


Figure A.4 – *Normalisation des données de comptage par la méthode de la médiane des ratios*

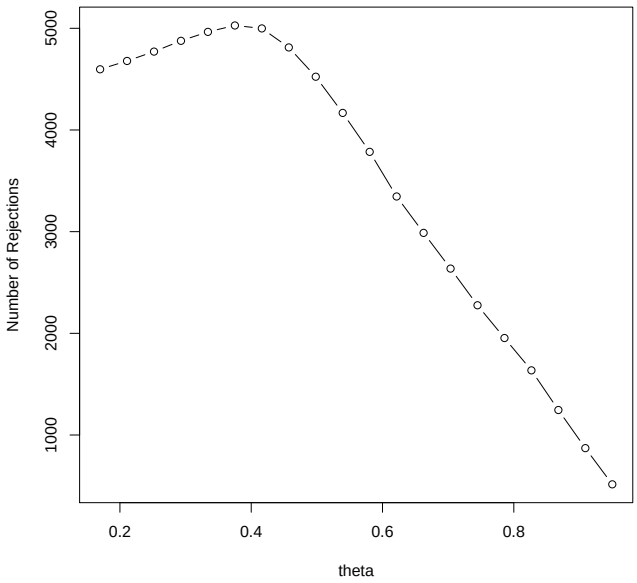


FIGURE A.5 – *Nombre de rejets de l'hypothèse nulle en fonction du pourcentage de gènes éliminés pour l'étude de l'impact de la LMMC. Le maximum est atteint en 0.375. Les gènes sont classés par ordre croissant de leur médiane d'expression dans tous les échantillons*

A.3.3 Résultats complémentaires

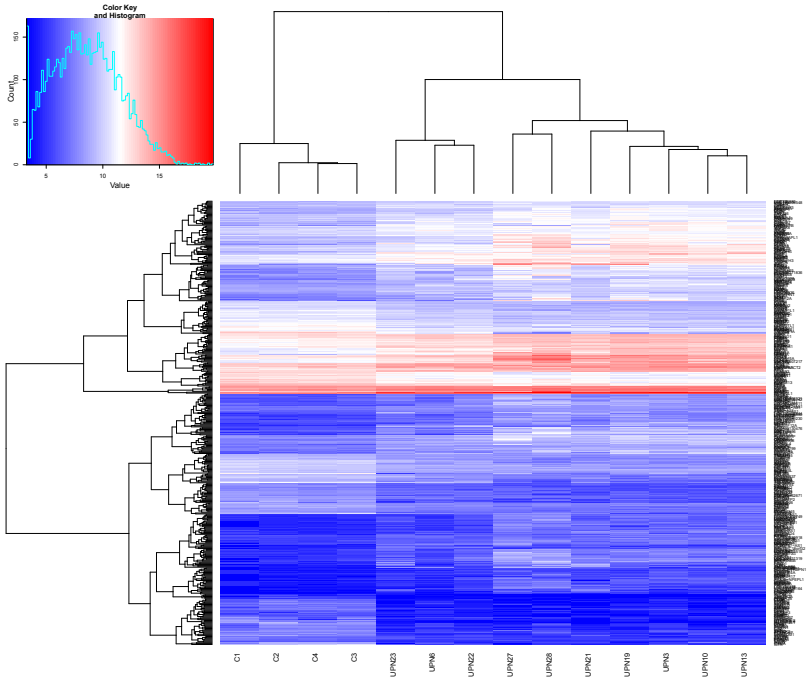


FIGURE A.6 – *Heatmap des 500 gènes les plus différentiellement exprimés dans l'étude de l'impact de la LMMC*

Gene	Exon(s)	Type	padj	$\Delta\psi$	Statut
ARID4B	18	Saut d'exon	$2.45\ 10^{-11}$	0.54	Validé
AMPD2	18	Saut d'exon	$6.1\ 10^{-11}$	0.51	Validé
TRIM33	18	Saut d'exon	$6.4\ 10^{-11}$	0.59	Validé
MLL	12	Saut d'exon	$1.5\ 10^{-10}$	0.56	Validé
TRIM21	4	Saut d'exon	$2.6\ 10^{-10}$	0.32	Validé
U2AF1	6	Saut d'exon	$8.4\ 10^{-9}$	0.22	Validé
ACSL1	14	Saut d'exon	$2\ 10^{-8}$	0.27	Validé
CTCF	3	Saut d'exon	$2.5\ 10^{-8}$	0.47	Validé
ACSL1	11	Saut d'exon	$1.1\ 10^{-7}$	0.28	Validé
RIT1	2	Saut d'exon	$1.0\ 10^{-4}$	0.32	Validé
PIM2	5	Saut d'exon	$1.7\ 10^{-3}$	0.46	Non validé
PKN2	20-21	Saut de plusieurs exons	$2.3\ 10^{-8}$	0.41	Validé
NOTCH2	22-23	Saut de plusieurs exons	$5.2\ 10^{-8}$	0.45	Validé
LUC7L3	2-4	Saut de plusieurs exons	$4.2\ 10^{-6}$	0.52	Validé
BID	3-4	Saut de plusieurs exons	$1.2\ 10^{-4}$	0.27	Validé
NUP98	20	Accepteur alternatif	$5.15\ 10^{-3}$	0.4	Validé
IRAK4	4	Rétention d'exon	$1.15\ 10^{-8}$	0.54	Validé

Table A.2 – *Épissages alternatifs détectés dans la leucémie myélomonocytaire chronique et testés par Q-PCR*

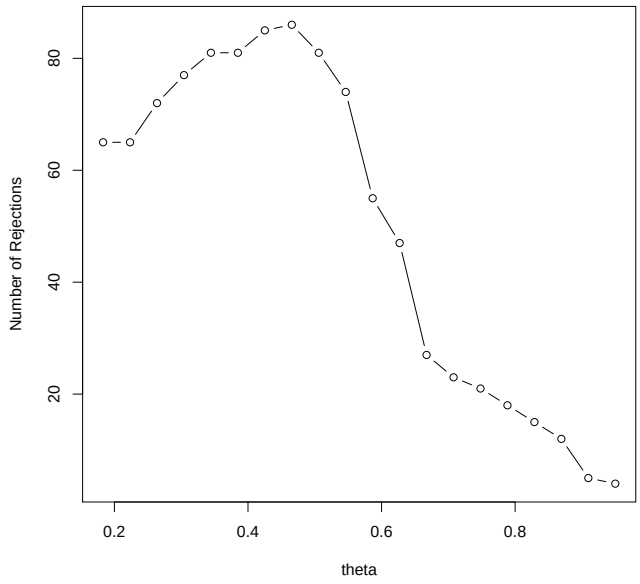


FIGURE A.7 – *Nombre de rejets de l'hypothèse nulle pour l'étude de l'impact de SRSF2 en fonction du pourcentage de gènes éliminés. Le maximum est atteint en 0.466. Les gènes sont classés par ordre croissant de leur médiane d'expression dans tous les échantillons*

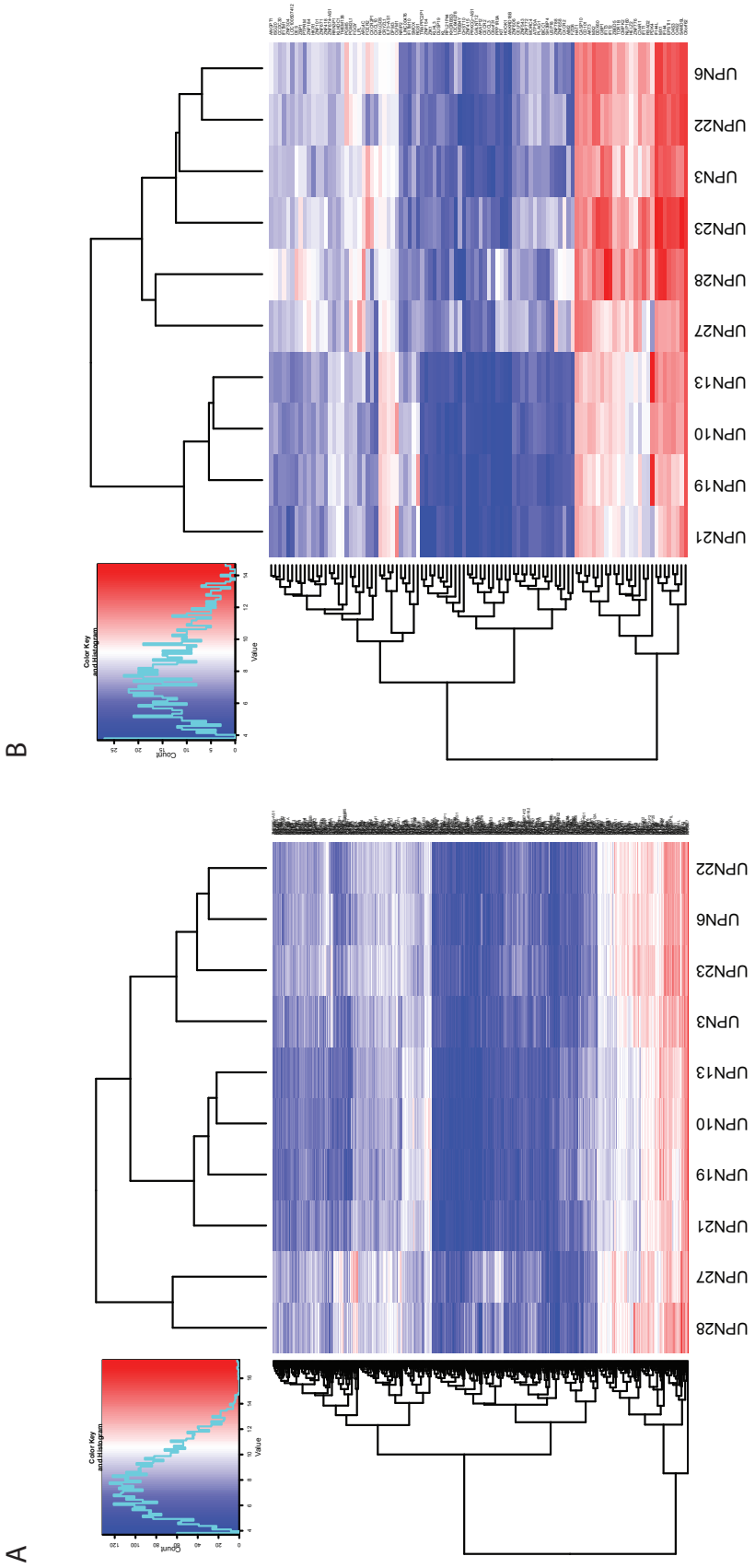


FIGURE A.8 – *Heatmap des 500 et 100 gènes les plus différemment exprimés dans l'étude de l'impact de la mutation de SRSF2*

Bibliographie

[1] Ahuja, N., Easwaran, H., Baylin, S. B., et al. (2014). Harnessing the potential of epigenetic therapy to target solid tumors. *The Journal of clinical investigation*, 124(124 (1)) :56–63. (Cité page 17.)

[2] Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., Mason, C. E., et al. (2012). methylkit : a comprehensive r package for the analysis of genome-wide dna methylation profiles. *Genome Biol*, 13(10) :R87. (Cité page 66.)

[3] Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013a). Signatures of mutational processes in human cancer. *Nature*, 500(7463) :415–421. (Cité pages 65, 74, 159 et 160.)

[4] Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., and Stratton, M. R. (2013b). Deciphering signatures of mutational processes operative in human cancer. *Cell reports*, 3(1) :246–259. (Cité pages 65 et 74.)

[5] Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*, 11(10) :R106. (Cité page 67.)

[6] Anders, S., Pyl, P. T., and Huber, W. (2014). Htseq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, page btu638. (Cité page 67.)

[7] Aucagne, R., Droin, N., Paggetti, J., Lagrange, B., Largeot, A., Hammann, A., Bataille, A., Martin, L., Yan, K.-P., Fenaux, P., et al. (2011). Transcription intermediary factor 1γ is a tumor suppressor in mouse and human chronic myelomonocytic leukemia. *The Journal of clinical investigation*, 121(6) :2361. (Cité page 32.)

[8] Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P., Stratton, M., et al. (2004). The cosmic (catalogue of somatic mutations in cancer) database and website. *British journal of cancer*, 91(2) :355–358. (Cité page 62.)

[9] Bednarek, A. K., Laflin, K. J., Daniel, R. L., Liao, Q., Hawkins, K. A., and Aldaz, C. M. (2000). Wwox, a novel ww domain-containing protein mapping to human chromosome 16q23. 3–24.1, a region frequently affected in breast cancer. *Cancer research*, 60(8) :2140–2145. (Cité page 161.)

[10] Bennett, J., Catovsky, D., Daniel, M., Flandrin, G., Galton, D., Gralnick, H., Sultan, C., and Cox, C. (1994). The chronic myeloid leukaemias : guidelines for distinguishing chronic granulocytic, atypical chronic myeloid, and chronic myelomonocytic leukaemia : Proposals by the french-american-british cooperative leukaemia group. *British journal of haematology*, 87(4) :746–754. (Cité page 7.)

[11] Bennett, J. M. (2000). World health organization classification of the acute leukemias and myelodysplastic syndrome. *International journal of hematology*, 72(2) :131–133. (Cité page 7.)

[12] Bennett, S. (2004). Solexa ltd. *Pharmacogenomics*, 5(4) :433–438. (Cité page 38.)

[13] Bennett, S., Barnes, C., Cox, A., Davies, L., and Brown, C. (2005). Toward the $1000 human genome. *Pharmacogenomics*, 6(4) :373–382. (Cité page 38.)

[14] Bentley, D. (2006). Whole-genome re-sequencing. *Current opinion in genetics & development*, 16(6) :545–552. (Cité page 38.)

[15] Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., Janoueix-Lerosey, I., Delattre, O., and Barillot, E. (2012). Control-freec : a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, 28(3) :423–425. (Cité page 63.)

[16] Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic : a flexible trimmer for illumina sequence data. *Bioinformatics*, page btu170. (Cité page 53.)

[17] Bos, J. L., Fearon, E. R., Hamilton, S. R., Verlaan-de Vries, M., van Boom, J. H., van der Eb, A. J., and Vogelstein, B. (1987). Prevalence of ras gene mutations in human colorectal cancers. *Nature*, 327(6120) :293–297. (Cité page 29.)

[18] Bos, J. L., Toksoz, D., Marshall, C. J., Verlaan-de Vries, M., Veeneman, G. H., van der Eb, A. J., van Boom, J. H., Janssen, J. W., and Steenvoorden, A. C. (1985). Amino-acid substitutions at codon 13 of the n-ras oncogene in human acute myeloid leukaemia. (Cité page 29.)

[19] Bowne, S., Humphries, M., Sullivan, L., Kenna, P., Tam, L., Kiang, A., Campbell, M., Weinstock, G., Koboldt, D., Ding, L., et al. (2011). A dominant mutation in rpe65 identified by whole-exome sequencing causes retinitis pigmentosa with choroidal involvement. *European journal of human genetics*, 19(10) :1074–1081. (Cité page 56.)

[20] Braun, T., Itzykson, R., Renneville, A., de Renzis, B., Dreyfus, F., Laribi, K., Bouabdallah, K., Vey, N., Toma, A., Recher, C., et al. (2011). Molecular predictors of response to decitabine in advanced chronic myelomonocytic leukemia : a phase 2 trial. *Blood*, 118(14) :3824–3831. (Cité pages 17, 33 et 148.)

[21] Bullard, J., Purdom, E., Hansen, K., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC bioinformatics*, 11(1) :94. (Cité page 67.)

[22] Busque, L., Patel, J. P., Figueroa, M. E., Vasanthakumar, A., Provost, S., Hamilou, Z., Mollica, L., Li, J., Viale, A., Heguy, A., et al. (2012). Recurrent somatic tet2 mutations in normal elderly individuals with clonal hematopoiesis. *Nature genetics*, 44(11) :1179–1181. (Cité pages 145 et 160.)

[23] Campbell, R. M., Tummino, P. J., et al. (2014). Cancer epigenetics drug discovery and development : the challenge of hitting the mark. *The Journal of clinical investigation*, 124(124 (1)) :64–69. (Cité page 19.)

[24] Chelala, C., Khan, A., and Lemoine, N. (2009). Snpnexus : a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics*, 25(5) :655–661. (Cité page 62.)

[25] Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., McGrath, S. D., Wendl, M. C., Zhang, Q., Locke, D. P., et al. (2009). Breakdancer : an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, 6(9) :677–681. (Cité page 65.)

[26] Chen, T., Hou, H., Chou, W., Tang, J., Kuo, Y., Chen, C., Tseng, M., Huang, C., Lai, Y., Chiang, Y., et al. (2014). Dynamics of asxl1 mutation and other associated genetic alterations during disease progression in patients with primary myelodysplastic syndrome. *Blood cancer journal*, 4(1) :e177. (Cité page 26.)

[27] Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J., Zhao, X., Carter, S. L., Russ, C., Nusbaum, C., Meyerson, M., and Lander, E. S. (2008). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature methods*, 6(1) :99–103. (Cité page 63.)

[28] Christman, J. K. (2002). 5-azacytidine and 5-aza-2'-deoxycytidine as inhibitors of dna methylation : mechanistic studies and their implications for cancer therapy. *Oncogene*, 21(35) :5483–5495. (Cité page 18.)

[29] Chun, S. and Fay, J. C. (2009). Identification of deleterious mutations within three human genomes. *Genome research*, 19(9) :1553–1561. (Cité page 62.)

[30] Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3) :213–219. (Cité page 56.)

[31] Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff : Snps in the genome of drosophila melanogaster strain w1118 ; iso-2 ; iso-3. *Fly*, 6(2) :80–92. (Cité page 62.)

[32] Craddock, C., Quek, L., Goardon, N., Freeman, S., Siddique, S., Raghavan, M., Aztberger, A., Schuh, A., Grimwade, D., Ivey, A., et al. (2013). Azacitidine fails to eradicate leukemic stem/progenitor cell populations in patients with acute myeloid leukemia and myelodysplasia. *Leukemia*, 27(5) :1028–1036. (Cité pages 162 et 163.)

[33] Czader, M. and Orazi, A. (2015). Acute myeloid leukemia and other types of disease progression in myeloproliferative neoplasms. *American journal of clinical pathology*, 144(2) :188–206. (Cité page 13.)

[34] Dalca, A. and Brudno, M. (2010). Genome variation discovery with high-throughput sequencing data. *Briefings in bioinformatics*, 11(1) :3–14. (Cité page 52.)

[35] Damm, F., Chesnais, V., Nagata, Y., Yoshida, K., Scourzic, L., Okuno, Y., Itzykson, R., Sanada, M., Shiraishi, Y., Gelsi-Boyer, V., et al. (2013). Bcor and bcorl1 mutations in myelodysplastic syndromes and related disorders. *Blood*, 122(18) :3169–3177. (Cité page 31.)

[36] David, M., Dzamba, M., Lister, D., Ilie, L., and Brudno, M. (2011). Shrimp2 : sensitive yet practical short read mapping. *Bioinformatics*, 27(7) :1011–1012. (Cité page 53.)

[37] Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using gerp++. *PLoS computational biology*, 6(12) :e1001025. (Cité page 62.)

[38] de la Mata, M., Alonso, C. R., Kadener, S., Fededa, J. P., Blaustein, M., Pelisch, F., Cramer, P., Bentley, D., and Kornblihtt, A. R. (2003). A slow rna polymerase ii affects alternative splicing in vivo. *Molecular cell*, 12(2) :525–532. (Cité page 31.)

[39] Delhommeau, F., Dupont, S., Valle, V. D., James, C., Trannoy, S., Masse, A., Kosmider, O., Le Couedic, J.-P., Robert, F., Alberdi, A., et al. (2009). Mutation in tet2 in myeloid cancers. *New England Journal of Medicine*, 360(22) :2289–2301. (Cité pages 23 et 25.)

[40] DePristo, M., Banks, E., Poplin, R., Garimella, K., Maguire, J., Hartl, C., Philippakis, A., Del Angel, G., Rivas, M., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, 43(5) :491–498. (Cité pages 52 et 55.)

[41] Derissen, E. J., Beijnen, J. H., and Schellens, J. H. (2013). Concise drug review : azacitidine and decitabine. *The oncologist*, 18(5) :619–624. (Cité page 18.)

[42] Devillier, R., Mansat-De Mas, V., Gelsi-Boyer, V., Demur, C., Murati, A., Corre, J., Prebet, T., Bertoli, S., Brecqueville, M., Arnoulet, C., et al. (2015). Role of asxl1 and tp53 mutations in the molecular classification and prognosis of acute myeloid leukemias with myelodysplasia-related changes. *Oncotarget*, 6(10) :8388–8396. (Cité page 26.)

[43] Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., et al. (2012). A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in Bioinformatics*. (Cité page 67.)

[44] Ding, J., Bashashati, A., Roth, A., Oloumi, A., Tse, K., Zeng, T., Haffari, G., Hirst, M., Marra, M., Condon, A., et al. (2012). Feature-based classifiers for somatic mutation detection in tumour–normal paired sequencing data. *Bioinformatics*, 28(2) :167–175. (Cité pages 55 et 56.)

[45] Ding, L., Wendl, M., Koboldt, D., and Mardis, E. (2010). Analysis of next-generation genomic data in cancer : accomplishments and challenges. *Human molecular genetics*, 19(R2) :R188–R196. (Cité page 52.)

[46] Durinck, S., Bullard, J., Spellman, P. T., and Dudoit, S. (2009). Genomegraphs : integrated genomic data visualization with r. *BMC bioinformatics*, 10(1) :2. (Cité page 67.)

[47] Fenaux, P., Mufti, G. J., Hellstrom-Lindberg, E., Santini, V., Finelli, C., Giagounidis, A., Schoch, R., Gattermann, N., Sanz, G., List, A., et al. (2009). Efficacy of azacitidine compared with that of conventional care regimens in the treatment of higher-risk myelodysplastic syndromes : a randomised, open-label, phase iii study. *The lancet oncology*, 10(3) :223–232. (Cité page 18.)

[48] Frankel, A. E., Lilly, M., Kreitman, R., Hogge, D., Beran, M., Freedman, M. H., Emanuel, P. D., McLain, C., Hall, P., Tagge, E., et al. (1998). Diphtheria toxin fused to granulocyte-macrophage colony-stimulating factor is toxic to blasts from patients with juvenile myelomonocytic leukemia and chronic myelomonocytic leukemia. *Blood*, 92(11) :4279–4286. (Cité pages 15 et 17.)

[49] Gambacorti-Passerini, C. B., Donadoni, C., Parmiani, A., Pirola, A., Redaelli, S., Signore, G., Piazza, V., Malcovati, L., Fontana, D., Spinelli, R., et al. (2015). Recurrent etnk1 mutations in atypical chronic myeloid leukemia. *Blood*, 125(3) :499–503. (Cité pages 32, 74 et 159.)

[50] Garrett-Bakelman, F. E., Sheridan, C. K., Kacmarczyk, T. J., Ishii, J., Betel, D., Alonso, A., Mason, C. E., Figueroa, M. E., and Melnick, A. M. (2015). Enhanced reduced representation bisulfite sequencing for assessment of dna methylation at base pair resolution. *Journal of visualized experiments : JoVE*, (96). (Cité page 66.)

[51] Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *Arxiv preprint arXiv :1207.3907*. (Cité page 55.)

[52] Ge, H., Liu, K., Juan, T., Fang, F., Newman, M., and Hoeck, W. (2011). Fusionmap : detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics*, 27(14) :1922–1928. (Cité page 70.)

[53] Gelsi-Boyer, V., Trouplin, V., Adélaïde, J., Bonansea, J., Cervera, N., Carbuccia, N., Lagarde, A., Prebet, T., Nezri, M., Sainty, D., et al. (2009). Mutations of polycomb-associated gene asxl1 in myelodysplastic syndromes and chronic myelomonocytic leukaemia. *British journal of haematology*, 145(6) :788–800. (Cité page 26.)

[54] Genovese, G., Kähler, A. K., Handsaker, R. E., Lindberg, J., Rose, S. A., Bakhoum, S. F., Chambert, K., Mick, E., Neale, B. M., Fromer, M., et al. (2014). Clonal hematopoiesis and blood-cancer risk inferred from blood dna sequence. *New England Journal of Medicine*, 371(26) :2477–2487. (Cité pages 145 et 160.)

[55] Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., et al. (2003). The international hapmap project. *Nature*, 426(6968) :789–796. (Cité page 61.)

[56] Gomez-Segui, I., Makishima, H., Jerez, A., Yoshida, K., Przychodzen, B., Miyano, S., Shiraishi, Y., Husseinzadeh, H., Guinta, K., Clemente, M., et al. (2013). Novel recurrent mutations in the ras-like gtp-binding gene rit1 in myeloid malignancies. *Leukemia*, 27(9) :1943. (Cité page 31.)

[57] Gondek, L. P., Haddad, A. S., O'Keefe, C. L., Tiu, R., Wlodarski, M. W., Sekeres, M. A., Theil, K. S., and Maciejewski, J. P. (2007). Detection of cryptic chromosomal lesions including acquired segmental uniparental disomy in advanced and low-risk myelodysplastic syndromes. *Experimental hematology*, 35(11) :1728–1738. (Cité page 21.)

[58] Goya, R., Sun, M., Morin, R., Leung, G., Ha, G., Wiegand, K., Senz, J., Crisan, A., Marra, M., Hirst, M., et al. (2010). Snvmix : predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, 26(6) :730–736. (Cité page 55.)

[59] Grada, A. and Weinbrecht, K. (2013). Next-generation sequencing : methodology and application. *Journal of Investigative Dermatology*, 133(8) :e11. (Cité page 42.)

[60] Grand, F. H., Hidalgo-Curtis, C. E., Ernst, T., Zoi, K., Zoi, C., McGuire, C., Kreil, S., Jones, A., Score, J., Metzgeroth, G., et al. (2009). Frequent cbl mutations associated with 11q acquired uniparental disomy in myeloproliferative neoplasms. *Blood*, 113(24) :6182–6192. (Cité page 21.)

[61] Graubert, T., Shen, D., Ding, L., Okeyo-Owuor, T., Lunn, C., Shao, J., Krysiak, K., Harris, C., Koboldt, D., Larson, D., et al. (2011). Recurrent mutations in the u2af1 splicing factor in myelodysplastic syndromes. *Nature genetics*. (Cité page 56.)

[62] Gualtieri, R. J., Emanuel, P. D., Zuckerman, K. S., Martin, G., Clark, S. C., Shadduck, R. K., Dracker, R. A., Akabutu, J., Nitschke, R., and Hetherington, M. L. (1989). Granulocyte-macrophage colony-stimulating factor is an endogenous regulator of cell proliferation in juvenile chronic myelogenous leukemia. *Blood*, 74(7) :2360–2367. (Cité page 15.)

[63] Guo, J. U., Su, Y., Zhong, C., Ming, G.-l., and Song, H. (2011). Hydroxylation of 5-methylcytosine by tet1 promotes active dna demethylation in the adult brain. *Cell*, 145(3) :423–434. (Cité page 25.)

[64] Hirai, H., Kobayashi, Y., Mano, H., Hagiwara, K., Maru, Y., Omine, M., Mizoguchi, H., Nishida, J., and Takaku, F. (1987). A point mutation at codon 13 of the n-ras oncogene in myelodysplastic syndrome. *Nature*, 327(6121) :430–432. (Cité page 29.)

[65] Hirata, H., Hinoda, Y., Shahryari, V., Deng, G., Nakajima, K., Tabatabai, Z. L., Ishii, N., and Dahiya, R. (2015). Long noncoding rna malat1 promotes aggressive renal cell carcinoma through ezh2 and interacts with mir-205. *Cancer research*, 75(7) :1322–1331. (Cité page 146.)

[66] Hu, L., Li, Z., Cheng, J., Rao, Q., Gong, W., Liu, M., Shi, Y. G., Zhu, J., Wang, P., and Xu, Y. (2013). Crystal structure of tet2-dna complex : insight into tet-mediated 5mc oxidation. *Cell*, 155(7) :1545–1555. (Cité page 25.)

[67] Hu, L., Wu, Y., Tan, D., Meng, H., Wang, K., Bai, Y., and Yang, K. (2015). Up-regulation of long noncoding rna malat1 contributes to proliferation and metastasis in esophageal squamous cell carcinoma. *J Exp Clin Cancer Res*, 34(7). (Cité page 146.)

[68] Ilagan, J. O., Ramakrishnan, A., Hayes, B., Murphy, M. E., Zebari, A. S., Bradley, P., and Bradley, R. K. (2015). U2af1 mutations alter splice site recognition in hematological malignancies. *Genome research*, 25(1) :14–26. (Cité page 33.)

[69] Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). De novo assembly and genotyping of variants using colored de bruijn graphs. *Nature Genetics*. (Cité page 55.)

[70] Itzykson, R., Droin, N., and Solary, É. (2012). Les progrès récents dans la leucémie myélomonocytaire chronique. *Hématologie*, 18(1) :24–36. (Cité page 1.)

[71] Itzykson, R., Kosmider, O., Renneville, A., Gelsi-Boyer, V., Meggendorfer, M., Morabito, M., Berthon, C., Adès, L., Fenaux, P., Beyne-Rauzy, O., et al. (2013a). Prognostic score including gene mutations in chronic myelomonocytic leukemia. *Journal of Clinical Oncology*. (Cité pages 22, 23, 26, 28, 30 et 31.)

[72] Itzykson, R., Kosmider, O., Renneville, A., Morabito, M., Preudhomme, C., Berthon, C., Adès, L., Fenaux, P., Platzbecker, U., Gagey, O., et al. (2013b). Clonal architecture of chronic myelomonocytic leukemias. *Blood*, 121(12) :2186–2198. (Cité pages 15, 32, 35 et 159.)

[73] Iyer, M. K., Chinnaiyan, A. M., and Maher, C. A. (2011). Chimerascan : a tool for identifying chimeric transcription in sequencing data. *Bioinformatics*, 27(20) :2903–2904. (Cité page 70.)

[74] Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P. V., Mar, B. G., Lindsley, R. C., Mermel, C. H., Burtt, N., Chavez, A., et al. (2014). Age-related clonal hematopoiesis associated with adverse outcomes. *New England Journal of Medicine*, 371(26) :2488–2498. (Cité pages 145 et 160.)

[75] James, C., Ugo, V., Le Couédic, J.-P., Staerk, J., Delhommeau, F., Lacout, C., Garçon, L., Raslova, H., Berger, R., Bennaceur-Griscelli, A., et al. (2005). A unique clonal jak2 mutation leading to constitutive signalling causes polycythaemia vera. *Nature*, 434(7037) :1144–1148. (Cité pages 13 et 30.)

[76] Janin, M., Mylonas, E., Saada, V., Micol, J.-B., Renneville, A., Quivoron, C., Koscielny, S., Scourzic, L., Forget, S., Pautas, C., et al. (2014). Serum 2-hydroxyglutarate production in idh1- and idh2-mutated de novo acute myeloid leukemia : a study by the acute leukemia french association group. *Journal of Clinical Oncology*, 32(4) :297–305. (Cité page 27.)

[77] Jankowska, A. M., Makishima, H., Tiu, R. V., Szpurka, H., Huang, Y., Traina, F., Visconte, V., Sugimoto, Y., Prince, C., O'Keefe, C., et al. (2011). Mutational spectrum analysis of chronic myelomonocytic leukemia includes genes associated with epigenetic regulation : Utx, ezh2, and dnmt3a. *Blood*, 118(14) :3932–3941. (Cité pages 25 et 26.)

[78] Jankowska, A. M., Szpurka, H., Tiu, R. V., Makishima, H., Afable, M., Huh, J., O'Keefe, C. L., Ganetzky, R., McDevitt, M. A., and Maciejewski, J. P. (2009). Loss of heterozygosity 4q24 and tet2 mutations associated with myelodysplastic/myeloproliferative neoplasms. *Blood*, 113(25) :6403–6410. (Cité page 21.)

[79] Jansen, M. P., Sas, L., Sieuwerts, A. M., Van Cauwenberghe, C., Ramirez-Ardila, D., Look, M., Ruigrok-Ritstier, K., Finetti, P., Bertucci, F., Timmermans, M. M., et al. (2015). Decreased expression of abat and stc2 hallmarks er-positive inflammatory breast cancer and endocrine therapy resistance in advanced disease. *Molecular oncology*, 9(6) :1218–1233. (Cité page 146.)

[80] Jia, P., Li, F., Xia, J., Chen, H., Ji, H., Pao, W., and Zhao, Z. (2012). Consensus rules in variant detection from next-generation sequencing data. *PloS one*, 7(6) :e38470. (Cité page 55.)

[81] Jiang, J., Guo, W., and Liang, X. (2014). Phenotypes, accumulation, and functions of myeloid-derived suppressor cells and associated treatment strategies in cancer patients. *Human immunology*, 75(11) :1128–1137. (Cité page 16.)

[82] Kaiser, J. (2008). A plan to capture human diversity in 1000 genomes. *Science*, 319(5862) :395–395. (Cité page 61.)

[83] Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A., et al. (2013). Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471) :333–339. (Cité page 159.)

[84] Kantarjian, H., Issa, J.-P. J., Rosenfeld, C. S., Bennett, J. M., Albitar, M., DiPersio, J., Klimek, V., Slack, J., de Castro, C., Ravandi, F., et al. (2006). Decitabine improves patient outcomes in myelodysplastic syndromes. *Cancer*, 106(8) :1794–1803. (Cité page 18.)

[85] Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010). Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nature methods*, 7(12) :1009–1015. (Cité page 69.)

[86] Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution : diversification, exon definition and function. *Nature Reviews Genetics*, 11(5) :345–355. (Cité page 34.)

[87] Khaled, Y. S., Ammori, B. J., and Elkord, E. (2013). Myeloid-derived suppressor cells in cancer : recent progress and prospects. *Immunology and cell biology*, 91(8) :493–502. (Cité page 16.)

[88] Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). Tophat2 : accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4) :R36. (Cité page 66.)

[89] Kim, E., Ilagan, J. O., Liang, Y., Daubner, G. M., Lee, S. C.-W., Ramakrishnan, A., Li, Y., Chung, Y. R., Micol, J.-B., Murphy, M. E., et al. (2015). Srsf2 mutations contribute to myelodysplasia by mutant-specific effects on exon recognition. *Cancer cell*, 27(5) :617–630. (Cité page 33.)

[90] Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D.-A., Mitterecker, A., Bodenhofer, U., and Hochreiter, S. (2012). cn. mops : mixture of poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Research*, 40(9) :e69–e69. (Cité page 63.)

[91] Knief, C. (2014). Analysis of plant microbe interactions in the era of next generation sequencing technologies. *Frontiers in plant science*, 5. (Cité page 37.)

[92] Ko, M., Huang, Y., Jankowska, A. M., Pape, U. J., Tahiliani, M., Bandukwala, H. S., An, J., Lamperti, E. D., Koh, K. P., Ganetzky, R., et al. (2010). Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant tet2. *Nature*, 468(7325) :839–843. (Cité page 25.)

[93] Koboldt, D., Chen, K., Wylie, T., Larson, D., McLellan, M., Mardis, E., Weinstock, G., Wilson, R., and Ding, L. (2009). Varscan : variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17) :2283–2285. (Cité page 55.)

[94] Koboldt, D., Zhang, Q., Larson, D., Shen, D., McLellan, M., Lin, L., Miller, C., Mardis, E., Ding, L., and Wilson, R. (2012). Varscan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3) :568–576. (Cité page 56.)

[95] Komeno, Y., Huang, Y.-J., Qiu, J., Lin, L., Xu, Y., Zhou, Y., Chen, L., Monterroza, D. D., Li, H., DeKelver, R. C., et al. (2015). Srsf2 is essential for hematopoiesis and its myelodysplastic syndromes-related mutations dysregulate alternative pre-mrna splicing. *Molecular and Cellular Biology*, pages MCB–00202. (Cité page 165.)

[96] Kon, A., Shih, L.-Y., Minamino, M., Sanada, M., Shiraishi, Y., Nagata, Y., Yoshida, K., Okuno, Y., Bando, M., Nakato, R., et al. (2013). Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. *Nature genetics*, 45(10) :1232–1237. (Cité pages 1, 23 et 28.)

[97] Krueger, F. and Andrews, S. R. (2011). Bismark : a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, 27(11) :1571–1572. (Cité page 66.)

[98] Kuo, M., Liang, D., Huang, C., Shih, Y., Wu, J., Lin, T., and Shih, L. (2009). Runx1 mutations are frequent in chronic myelomonocytic leukemia and mutations at the c-terminal region might predict acute myeloid leukemia transformation. *Leukemia*, 23(8) :1426–1431. (Cité page 31.)

[99] Lai, D. and Ha, G. (2012). Hmmcopy : A package for bias-free copy number estimation and robust cna detection in tumour samples from wgs hts data. (Cité page 63.)

[100] Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., and Maglott, D. R. (2014). Clinvar : public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*, 42(D1) :D980–D985. (Cité page 62.)

[101] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4) :357–359. (Cité page 53.)

[102] Langmead, B., Trapnell, C., Pop, M., Salzberg, S., et al. (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3) :R25. (Cité page 53.)

[103] Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., Ley, T. J., Mardis, E. R., Wilson, R. K., and Ding, L. (2012). Somaticsniper : identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3) :311–317. (Cité pages 56, 57, 58 et 59.)

[104] Lee, J.-H., Choi, Y., Kim, S.-D., Kim, D.-Y., Lee, J.-H., Lee, K.-H., Lee, S.-M., Lee, W.-S., and Joo, Y.-D. (2015). Clinical outcome after failure of hypomethylating therapy for myelodysplastic syndrome. *European journal of haematology*, 94(6) :546–553. (Cité page 17.)

[105] Lee, P. and Shatkay, H. (2008). F-snp : computationally predicted functional snps for disease association studies. *Nucleic acids research*, 36(suppl 1) :D820–D824. (Cité page 62.)

[106] Levine, R. L., Loriaux, M., Huntly, B. J., Loh, M. L., Beran, M., Stoffregen, E., Berger, R., Clark, J. J., Willis, S. G., Nguyen, K. T., et al. (2005). The jak2v617f activating mutation occurs in chronic myelomonocytic leukemia and acute myeloid leukemia, but not in acute lymphoblastic leukemia or chronic lymphocytic leukemia. *Blood*, 106(10) :3377–3379. (Cité page 30.)

[107] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14) :1754–1760. (Cité page 53.)

[108] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., et al. (2009a). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16) :2078–2079. (Cité page 55.)

[109] Li, H., Ruan, J., and Durbin, R. (2008a). Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11) :1851–1858. (Cité pages 53 et 55.)

[110] Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., and Wang, J. (2009b). Snp detection for massively parallel whole-genome resequencing. *Genome research*, 19(6) :1124–1132. (Cité page 55.)

[111] Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008b). Soap : short oligonucleotide alignment program. *Bioinformatics*, 24(5) :713–714. (Cité page 53.)

[112] Li, R., Yu, C., Li, Y., Lam, T., Yiu, S., Kristiansen, K., and Wang, J. (2009c). Soap2 : an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15) :1966–1967. (Cité page 53.)

[113] Li, Y., Chien, J., Smith, D. I., and Ma, J. (2011). Fusionhunter : identifying fusion transcripts in cancer using paired-end rna-seq. *Bioinformatics*, 27(12) :1708–1710. (Cité page 70.)

[114] Lister, R. and Ecker, J. R. (2009). Finding the fifth base : genome-wide sequencing of cytosine methylation. *Genome research*, 19(6) :959–966. (Cité page 66.)

[115] Liu, C., Wong, T., Wu, E., Luo, R., Yiu, S., Li, Y., Wang, B., Yu, C., Chu, X., Zhao, K., et al. (2012). Soap3 : Ultra-fast gpu-based parallel alignment tool for short reads. *Bioinformatics*. (Cité page 53.)

[116] Liu, E., Hjelle, B., Morgan, R., Hecht, F., and Bishop, J. M. (1987). Mutations of the kirsten-ras proto-oncogene in human preleukaemia. *Nature*, 330(6144) :186–188. (Cité page 29.)

[117] Lorenzo, F., Nishii, K., Monma, F., Kuwagata, S., Usui, E., and Shiku, H. (2006). Mutational analysis of the kit gene in myelodysplastic syndrome (mds) and mds-derived leukemia. *Leukemia research*, 30(10) :1235–1239. (Cité page 30.)

[118] Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, 15(12) :550. (Cité pages 68 et 69.)

[119] Love, M. I., Myšicková, A., Sun, R., Kalscheuer, V., Vingron, M., and Haas, S. A. (2011). Modeling read counts for cnv detection in exome sequencing data. *Statistical applications in genetics and molecular biology*, 10(1) :52. (Cité page 63.)

[120] Lübbert, M., Suciu, S., Baila, L., Rüter, B. H., Platzbecker, U., Giagounidis, A., Selleslag, D., Labar, B., Germing, U., Salih, H. R., et al. (2011). Low-dose decitabine versus best supportive care in elderly patients with intermediate-or high-risk myelodysplastic syndrome (mds) ineligible for intensive chemotherapy : final results of the randomized phase iii study of the european organisation for research and treatment of cancer leukemia group and the german mds study group. *Journal of clinical oncology*, pages JCO–2010. (Cité pages 17 et 18.)

[121] Luo, C., Tsementzi, D., Kyrpides, N., Read, T., and Konstantinidis, K. (2012). Direct comparisons of illumina vs. roche 454 sequencing technologies on the same microbial community dna sample. *PloS one*, 7(2) :e30087. (Cité page 38.)

[122] Ma, K.-x., Wang, H.-j., Li, X.-r., Li, T., Su, G., Yang, P., and Wu, J.-w. (2015). Long noncoding rna malat1 associates with the malignant status and poor prognosis in glioma. *Tumor Biology*, 36(5) :3355–3359. (Cité page 146.)

[123] Makarov, V., O'Grady, T., Cai, G., Lihm, J., Buxbaum, J., and Yoon, S. (2012). Anntools : a comprehensive and versatile annotation toolkit for genomic variants. *Bioinformatics*, 28(5) :724–725. (Cité page 62.)

[124] Marcucci, G., Maharry, K., Wu, Y.-Z., Radmacher, M. D., Mrózek, K., Margeson, D., Holland, K. B., Whitman, S. P., Becker, H., Schwind, S., et al. (2010). Idh1 and idh2 gene mutations identify novel molecular subsets within de novo cytogenetically normal acute myeloid leukemia : a cancer and leukemia group b study. *Journal of Clinical Oncology*, 28(14) :2348–2355. (Cité page 27.)

[125] Mardis, E. (2008). Next-generation dna sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9 :387–402. (Cité page 38.)

[126] Mardis, E., Ding, L., Dooling, D., Larson, D., McLellan, M., Chen, K., Koboldt, D., Fulton, R., Delehaunty, K., McGrath, S., et al. (2009). Recurring mutations found by sequencing an acute myeloid leukemia genome. *New England Journal of Medicine*, 361(11) :1058–1066. (Cité page 49.)

[127] Margulies, M., Egholm, M., Altman, W., Attiya, S., Bader, J., Bemben, L., Berka, J., Braverman, M., Chen, Y., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057) :376–380. (Cité page 38.)

[128] Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1) :pp–10. (Cité page 53.)

[129] Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D. C., Fullam, A., Alexandrov, L. B., Tubio, J. M., et al. (2015). High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348(6237) :880–886. (Cité page 145.)

[130] Matsushita, H., Vesely, M., Koboldt, D., Rickert, C., Uppaluri, R., Magrini, V., Arthur, C., White, J., Chen, Y., Shea, L., et al. (2012). Cancer exome analysis reveals a t-cell-dependent mechanism of cancer immunoediting. *Nature*, 482(7385) :400–404. (Cité page 56.)

[131] Maxson, J. E., Gotlib, J., Pollyea, D. A., Fleischman, A. G., Agarwal, A., Eide, C. A., Bottomly, D., Wilmot, B., McWeeney, S. K., Tognon, C. E., et al. (2013). Oncogenic csf3r mutations in chronic neutrophilic leukemia and atypical cml. *New England Journal of Medicine*, 368(19) :1781–1790. (Cité page 14.)

[132] McCabe, M. T., Brandes, J. C., and Vertino, P. M. (2009). Cancer dna methylation : molecular mechanisms and clinical implications. *Clinical Cancer Research*, 15(12) :3927–3937. (Cité page 18.)

[133] McCabe, M. T., Ott, H. M., Ganji, G., Korenchuk, S., Thompson, C., Van Aller, G. S., Liu, Y., Graves, A. P., Diaz, E., LaFrance, L. V., et al. (2012). Ezh2 inhibition as a therapeutic strategy for lymphoma with ezh2-activating mutations. *Nature*, 492(7427) :108–112. (Cité page 19.)

[134] McFarland, C. D., Korolev, K. S., Kryukov, G. V., Sunyaev, S. R., and Mirny, L. A. (2013). Impact of deleterious passenger mutations on cancer progression. *Proceedings of the National Academy of Sciences*, 110(8) :2910–2915. (Cité pages 161 et 164.)

[135] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit : a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9) :1297–1303. (Cité page 55.)

[136] McPherson, A., Hormozdiari, F., Zayed, A., Giuliany, R., Ha, G., Sun, M. G., Griffith, M., Moussavi, A. H., Senz, J., Melnyk, N., et al. (2011). defuse : an algorithm for gene fusion discovery in tumor rna-seq data. (Cité page 70.)

[137] Medina, I., De Maria, A., Bleda, M., Salavert, F., Alonso, R., Gonzalez, C. Y., and Dopazo, J. (2012). Variant : Command line, web service and web interface for fast and accurate functional characterization of variants found by next-generation sequencing. *Nucleic acids research*, 40(W1) :W54–W58. (Cité page 62.)

[138] Meggendorfer, M., Roller, A., Haferlach, T., Eder, C., Dicker, F., Grossmann, V., Kohlmann, A., Alpermann, T., Yoshida, K., Ogawa, S., et al. (2012). Srsf2 mutations in 275 cases with chronic myelomonocytic leukemia (cmml). *Blood*, 120(15) :3080–3088. (Cité page 28.)

[139] Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis. *Nucleic acids research*, 33(18) :5868–5877. (Cité page 66.)

[140] Meldi, K., Qin, T., Buchi, F., Droin, N., Sotzen, J., Micol, J.-B., Selimoglu-Buet, D., Masala, E., Allione, B., Gioia, D., et al. (2015). Specific molecular signatures predict decitabine response in chronic myelomonocytic leukemia. *The Journal of clinical investigation*, 125(125 (5)) :0–0. (Cité page 17.)

[141] Metzker, M. (2009). Sequencing technologies-the next generation. *Nature Reviews Genetics*, 11(1) :31–46. (Cité page 38.)

[142] Micol, J.-B., Duployez, N., Boissel, N., Petit, A., Geffroy, S., Nibourel, O., Lacombe, C., Lapillonne, H., Etancelin, P., Figeac, M., et al. (2014). Frequent asxl2 mutations in acute myeloid leukemia patients with t (8 ; 21)/runx1-runx1t1 chromosomal translocations. *Blood*, 124(9) :1445–1449. (Cité pages 74 et 159.)

[143] Mitelman, F., Nilsson, P., Levan, G., and Brandt, L. (1976). Non-random chromosome changes in acute myeloid leukemia. chromosome banding examination of 30 cases at diagnosis. *International Journal of Cancer*, 18(1) :31–38. (Cité page 22.)

[144] Mohamedali, A. M., Smith, A. E., Gaken, J., Lea, N. C., Mian, S. A., Westwood, N. B., Strupp, C., Gattermann, N., Germing, U., and Mufti, G. J. (2009). Novel tet2 mutations associated with upd4q24 in myelodysplastic syndrome. *Journal of Clinical Oncology*, 27(24) :4002–4006. (Cité page 21.)

[145] Morozova, O. and Marra, M. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5) :255–264. (Cité page 38.)

[146] Nangalia, J., Massie, C. E., Baxter, E. J., Nice, F. L., Gundem, G., Wedge, D. C., Avezov, E., Li, J., Kollmann, K., Kent, D. G., et al. (2013). Somatic calr mutations in myeloproliferative neoplasms with nonmutated jak2. *New England Journal of Medicine*, 369(25) :2391–2405. (Cité page 13.)

[147] Ng, P. and Henikoff, S. (2003). Sift : Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13) :3812–3814. (Cité page 62.)

[148] Ng, S., Buckingham, K., Lee, C., Bigham, A., Tabor, H., Dent, K., Huff, C., Shannon, P., Jabs, E., Nickerson, D., et al. (2009). Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics*, 42(1) :30–35. (Cité page 49.)

[149] Nielsen, R., Paul, J., Albrechtsen, A., and Song, Y. (2011). Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6) :443–451. (Cité page 52.)

[150] Obenchain, V., Lawrence, M., Carey, V., Gogarten, S., Shannon, P., and Morgan, M. (2014). Variantannotation : a bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*, 30(14) :2076–2078. (Cité page 62.)

[151] Oh, S. T., Simonds, E. F., Jones, C., Hale, M. B., Goltsev, Y., Gibbs Jr, K. D., Merker, J. D., Zehnder, J. L., Nolan, G. P., and Gotlib, J. (2010). Novel mutations in the inhibitory adaptor protein lnk drive jak-stat signaling in patients with myeloproliferative neoplasms. *Blood*, 116(6) :988–992. (Cité page 30.)

[152] Ortmann, C. A., Kent, D. G., Nangalia, J., Silber, Y., Wedge, D. C., Grinfeld, J., Baxter, E. J., Massie, C. E., Papaemmanuil, E., Menon, S., et al. (2015). Effect of mutation order on myeloproliferative neoplasms. *New England Journal of Medicine*, 372(7) :601–612. (Cité page 159.)

[153] Padron, E., Painter, J. S., Kunigal, S., Mailloux, A. W., McGraw, K., McDaniel, J. M., Kim, E., Bebbington, C., Baer, M., Yarranton, G., et al. (2013). Gm-csf–dependent pstat5 sensitivity is a feature with therapeutic potential in chronic myelomonocytic leukemia. *Blood*, 121(25) :5068–5077. (Cité pages 16 et 17.)

[154] Pardanani, A., Lasho, T., Finke, C., Mai, M., McClure, R., and Tefferi, A. (2010). Idh1 and idh2 mutation analysis in chronic-and blast-phase myeloproliferative neoplasms. *Leukemia*, 24(6) :1146–1151. (Cité page 27.)

[155] Pasquali, F., Bernasconi, P., Casalone, R., Fraccaro, M., Bernasconi, C., Lazzarino, M., Morra, E., Alessandrino, E., Marchi, M., and Sanger, R. (1982). Pathogenetic significance of "pure" monosomy 7 in myeloproliferative disorders. analysis of 14 cases. *Human genetics*, 62(1) :40–51. (Cité page 22.)

[156] Patnaik, M., Wassie, E., Padron, E., Onida, F., Itzykson, R., Lasho, T., Kosmider, O., Finke, C., Hanson, C., Ketterling, R., et al. (2015). Chronic myelomonocytic leukemia in younger patients : molecular and cytogenetic predictors of survival and treatment outcome. *Blood cancer journal*, 4(1) :e270. (Cité page 9.)

[157] Pattnaik, S., Vaidyanathan, S., Pooja, D. G., Deepak, S., and Panda, B. (2012). Customisation of the exome data analysis pipeline using a combinatorial approach. *PLoS One*, 7(1) :e30080. (Cité page 52.)

[158] Pepper, J. W., Sprouffske, K., and Maley, C. C. (2007). Animal cell differentiation patterns suppress somatic evolution. *PLoS Comput Biol*, 3(12) :e250. (Cité page 10.)

[159] Perié, L., Hodgkin, P. D., Naik, S. H., Schumacher, T. N., de Boer, R. J., and Duffy, K. R. (2014). Determining lineage pathways from cellular barcoding experiments. *Cell reports*, 6(4) :617–624. (Cité page 12.)

[160] Petit, P., Alexander, M., and Fondu, P. (1973). Monosomy 7 in erythroleukaemia. *The Lancet*, 302(7841) :1326–1327. (Cité page 22.)

[161] Philippe, N., Salson, M., Commes, T., and Rivals, E. (2013). Crac : an integrated approach to the analysis of rna-seq reads. *Genome biology*, 14(3) :R30. (Cité pages 66 et 70.)

[162] Pilati, C., Letouzé, E., Nault, J.-C., Imbeaud, S., Boulai, A., Calderaro, J., Poussin, K., Franconi, A., Couchy, G., Morcrette, G., et al. (2014). Genomic profiling of hepatocellular adenomas reveals recurrent frk-activating mutations and the mechanisms of malignant transformation. *Cancer Cell*, 25(4) :428–441. (Cité page 160.)

[163] Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.-L., Chen, C., Zhai, Y., et al. (1998). High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nature genetics*, 20(2) :207–211. (Cité page 63.)

[164] Pollard, K., Hubisz, M., Rosenbloom, K., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1) :110–121. (Cité page 62.)

[165] Przychodzen, B., Jerez, A., Guinta, K., Sekeres, M. A., Padgett, R., Maciejewski, J. P., and Makishima, H. (2013). Patterns of missplicing due to somatic u2af1 mutations in myeloid neoplasms. *Blood*, 122(6) :999–1006. (Cité page 33.)

[166] Puente, X. S., Beà, S., Valdés-Mas, R., Villamor, N., Gutiérrez-Abril, J., Martín-Subero, J. I., Munar, M., Rubio-Pérez, C., Jares, P., Aymerich, M., et al. (2015). Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*. (Cité page 160.)

[167] Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous snps : server and survey. *Nucleic acids research*, 30(17) :3894–3900. (Cité page 62.)

[168] Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., and Korbel, J. O. (2012). Delly : structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18) :i333–i339. (Cité page 65.)

[169] Raza, A. and Galili, N. (2012). The genetic basis of phenotypic heterogeneity in myelodysplastic syndromes. *Nature Reviews Cancer*, 12(12) :849–859. (Cité page 30.)

[170] Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biol*, 11(3) :R25. (Cité page 67.)

[171] Roth, A., Morin, R., Ding, J., Crisan, A., Ha, G., Giuliany, R., Bashashati, A., Hirst, M., Turashvili, G., Oloumi, A., et al. (2012). Jointsnvmix : A probabilistic model for accurate detection of somatic mutations in normal/tumour paired next generation sequencing data. *Bioinformatics*. (Cité page 55.)

[172] Rowley, J. D. (1980). Chromosome abnormalities in acute lymphoblastic leukemia. *Cancer Genetics and Cytogenetics*, 1(3) :263–271. (Cité page 22.)

[173] Rumble, S., Lacroute, P., Dalca, A., Fiume, M., Sidow, A., and Brudno, M. (2009). Shrimp : accurate mapping of short color-space reads. *PLoS computational biology*, 5(5) :e1000386. (Cité page 53.)

[174] Sanada, M., Suzuki, T., Shih, L.-Y., Otsu, M., Kato, M., Yamazaki, S., Tamura, A., Honda, H., Sakata-Yanagimoto, M., Kumano, K., et al. (2009). Gain-of-function of mutated c-cbl tumour suppressor in myeloid neoplasms. *Nature*, 460(7257) :904–908. (Cité page 29.)

[175] Sathirapongsasuti, J. F., Lee, H., Horst, B. A., Brunner, G., Cochran, A. J., Binder, S., Quackenbush, J., and Nelson, S. F. (2011). Exome sequencing-based copy-number variation and loss of heterozygosity detection : Exomecnv. *Bioinformatics*, 27(19) :2648–2654. (Cité page 63.)

[176] Saunders, C., Wong, W., Swamy, S., Becq, J., Murray, L., and Cheetham, R. (2012). Strelka : accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, 28(14) :1811–1817. (Cité pages 56, 59 et 60.)

[177] Sboner, A., Habegger, L., Pflueger, D., Terry, S., Chen, D. Z., Rozowsky, J. S., Tewari, A. K., Kitabayashi, N., Moss, B. J., Chee, M. S., et al. (2010). Fusionseq : a modular framework for finding gene fusions by analyzing paired-end rna-sequencing data. *Genome Biol*, 11(10) :R104. (Cité page 70.)

[178] Schwarz, J., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010). Mutationtaster evaluates disease-causing potential of sequence alterations. *Nature methods*, 7(8) :575–576. (Cité page 62.)

[179] Selimoglu-Buet, D., Wagner-Ballon, O., Saada, V., Bardet, V., Itzykson, R., Bencheikh, L., Morabito, M., Met, E., Debord, C., Benayoun, E., et al. (2015). Characteristic repartition of monocyte subsets as a diagnostic signature of chronic myelomonocytic leukemia. *Blood*, pages blood–2015. (Cité page 9.)

[180] Server, E. V. (2012). Nhlbi exome sequencing project (esp). *Seattle, WA*. (Cité page 61.)

[181] Shen, Q., Ouyang, J., Tang, G., Jabbour, E. J., Garcia-Manero, G., Routbort, M., Konoplev, S., Bueso-Ramos, C., Medeiros, L. J., Jorgensen, J. L., et al. (2015). Flow cytometry immunophenotypic findings in chronic myelomonocytic leukemia and its utility in monitoring treatment response. *European journal of haematology*. (Cité page 163.)

[182] Sherry, S., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E., and Sirotkin, K. (2001). dbsnp : the ncbi database of genetic variation. *Nucleic acids research*, 29(1) :308–311. (Cité page 61.)

[183] Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8) :1034–1050. (Cité page 62.)

[184] Silverman, L. R., Demakos, E. P., Peterson, B. L., Kornblith, A. B., Holland, J. C., Odchimar-Reissig, R., Stone, R. M., Nelson, D., Powell, B. L., DeCastro, C. M., et al. (2002). Randomized controlled trial of azacitidine in patients with the myelodysplastic syndrome : a study of the cancer and leukemia group b. *Journal of Clinical oncology*, 20(10) :2429–2440. (Cité page 18.)

[185] Silverman, L. R., Fenaux, P., Mufti, G. J., Santini, V., Hellström-Lindberg, E., Gattermann, N., Sanz, G., List, A. F., Gore, S. D., and Seymour, J. F. (2011). Continued azacitidine therapy beyond time of first response improves quality of response in patients with higher-risk myelodysplastic syndromes. *Cancer*, 117(12) :2697–2702. (Cité page 17.)

[186] Sim, N., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P. (2012). Sift web server : predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*, 40(W1) :W452–W457. (Cité page 62.)

[187] Singh, H., Lane, A., Correll, M., Przychodzen, B., Sykes, D., Stone, R., Ballen, K., Amrein, P., Maciejewski, J., and Attar, E. (2013). Putative rna-splicing gene luc7l2 on 7q34 represents a candidate gene in pathogenesis of myeloid malignancies. *Blood cancer journal*, 3(5) :e117. (Cité pages 28, 74 et 159.)

[188] Steensma, D. P., Bejar, R., Jaiswal, S., Lindsley, R. C., Sekeres, M. A., Hasserjian, R. P., and Ebert, B. L. (2015). Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood*, pages blood–2015. (Cité page 160.)

[189] Strober, S. (1984). Natural suppressor (ns) cells, neonatal tolerance, and total lymphoid irradiation : exploring obscure relationships. *Annual review of immunology*, 2(1) :219–237. (Cité page 16.)

[190] Such, E., Cervera, J., Costa, D., Solé, F., Vallespí, T., Luño, E., Collado, R., Calasanz, M. J., Hernández-Rivas, J. M., Cigudosa, J. C., et al. (2011). Cytogenetic risk stratification in chronic myelomonocytic leukemia. *Haematologica*, 96(3) :375–383. (Cité pages 8 et 22.)

[191] Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. *Cell*, 156(6) :1324–1335. (Cité page 162.)

[192] Tang, G., Fu, B., Hu, S., Lu, X., Tang, Z., Li, S., Jabbar, K., Khoury, J. D., Medeiros, L. J., and Wang, S. A. (2015). Prognostic impact of acquisition of cytogenetic abnormalities during the course of chronic myelomonocytic leukemia. *American journal of hematology*. (Cité pages 8 et 22.)

[193] Tefferi, A., Lasho, T., Abdel-Wahab, O., Guglielmelli, P., Patel, J., Caramazza, D., Pieri, L., Finke, C., Kilpivaara, O., Wadleigh, M., et al. (2010). Idh1 and idh2 mutation studies in 1473 patients with chronic-, fibrotic-or blast-phase essential thrombocythemia, polycythemia vera or myelofibrosis. *Leukemia*, 24(7) :1302–1309. (Cité page 27.)

[194] Thoennissen, N. H., Lasho, T., Thoennissen, G. B., Ogawa, S., Tefferi, A., and Koeffler, H. P. (2011). Novel cux1 missense mutation in association with 7q- at leukemic transformation of mpn. *American journal of hematology*, 86(8) :703–705. (Cité page 31.)

[195] Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature biotechnology*, 31(1) :46–53. (Cité page 69.)

[196] Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). Tophat : discovering splice junctions with rna-seq. *Bioinformatics*, 25(9) :1105–1111. (Cité page 66.)

[197] Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, 7(3) :562–578. (Cité page 67.)

[198] Treppendahl, M. B., Kristensen, L. S., Grønbæk, K., et al. (2014). Predicting response to epigenetic therapy. *The Journal of clinical investigation*, 124(124 (1)) :47–55. (Cité pages 17 et 18.)

[199] Trikha, P. and Carson, W. E. (2014). Signaling pathways involved in mdsc regulation. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1846(1) :55–65. (Cité page 16.)

[200] Vardiman, J. W., Thiele, J., Arber, D. A., Brunning, R. D., Borowitz, M. J., Porwit, A., Harris, N. L., Le Beau, M. M., Hellström-Lindberg, E., Tefferi, A., et al. (2009). The 2008 revision of the world health organization (who) classification of myeloid neoplasms and acute leukemia : rationale and important changes. *Blood*, 114(5) :937–951. (Cité page 7.)

[201] Vinagre, J., Almeida, A., Pópulo, H., Batista, R., Lyra, J., Pinto, V., Coelho, R., Celestino, R., Prazeres, H., Lima, L., et al. (2013). Frequency of tert promoter mutations in human cancers. *Nature communications*, 4. (Cité page 160.)

[202] Voso, M. T., Santini, V., Fabiani, E., Fianchi, L., Criscuolo, M., Falconi, G., Guidi, F., Hohaus, S., and Leone, G. (2014). Why methylation is not a marker predictive of response to hypomethylating agents. *haematologica*, 99(4) :613–619. (Cité page 18.)

[203] Wang, F., Travins, J., DeLaBarre, B., Penard-Lacronique, V., Schalm, S., Hansen, E., Straley, K., Kernytsky, A., Liu, W., Gliser, C., et al. (2013a). Targeted inhibition of mutant idh2 in leukemia cells induces cellular differentiation. *Science*, 340(6132) :622–626. (Cité page 19.)

[204] Wang, K., Li, M., and Hakonarson, H. (2010). Annovar : functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16) :e164–e164. (Cité page 62.)

[205] Wang, Q., Jia, P., Li, F., Chen, H., Ji, H., Hucks, D., Dahlman, K. B., Pao, W., Zhao, Z., et al. (2013b). Detecting somatic point mutations in cancer genome sequencing data : a comparison of mutation callers. *Genome Med*, 5(10) :91. (Cité page 56.)

[206] Wassie, E. A., Itzykson, R., Lasho, T. L., Kosmider, O., Finke, C. M., Hanson, C. A., Ketterling, R. P., Solary, E., Tefferi, A., and Patnaik, M. M. (2014). Molecular and prognostic correlates of cytogenetic abnormalities in chronic myelomonocytic leukemia : a mayo clinic-french consortium study. *American journal of hematology*, 89(12) :1111–1115. (Cité page 8.)

[207] Watson, I. R., Takahashi, K., Futreal, P. A., and Chin, L. (2013). Emerging patterns of somatic mutations in cancer. *Nature Reviews Genetics*, 14(10) :703–718. (Cité page 25.)

[208] Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature genetics*, 46(11) :1160–1165. (Cité page 159.)

[209] Xie, M., Lu, C., Wang, J., McLellan, M. D., Johnson, K. J., Wendl, M. C., McMichael, J. F., Schmidt, H. K., Yellapantula, V., Miller, C. A., et al. (2014). Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nature medicine*, 20(12) :1472–1478. (Cité pages 145 et 160.)

[210] Yan, H., Parsons, D. W., Jin, G., McLendon, R., Rasheed, B. A., Yuan, W., Kos, I., Batinic-Haberle, I., Jones, S., Riggins, G. J., et al. (2009). Idh1 and idh2 mutations in gliomas. *New England Journal of Medicine*, 360(8) :765–773. (Cité pages 26 et 27.)

[211] Yandell, M., Huff, C., Hu, H., Singleton, M., Moore, B., Xing, J., Jorde, L., and Reese, M. (2011). A probabilistic disease-gene finder for personal genomes. *Genome research*, 21(9) :1529–1542. (Cité page 62.)

[212] Yang, H., Ye, D., Guan, K.-L., and Xiong, Y. (2012). Idh1 and idh2 mutations in tumorigenesis : mechanistic insights and clinical perspectives. *Clinical Cancer Research*, 18(20) :5562–5571. (Cité page 27.)

[213] Ye, K., Schulz, M. H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel : a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21) :2865–2871. (Cité page 65.)

[214] Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome research*, 19(9) :1586–1592. (Cité page 63.)

[215] You, N., Murillo, G., Su, X., Zeng, X., Xu, J., Ning, K., Zhang, S., Zhu, J., and Cui, X. (2012). Snp calling using genotype model selection on high-throughput sequencing data. *Bioinformatics*, 28(5) :643–650. (Cité pages 52 et 55.)

[216] Yunis, J. J., Brunning, R. D., Howe, R. B., and Lobell, M. (1984). High-resolution chromosomes as an independent prognostic indicator in adult acute nonlymphocytic leukemia. *New England Journal of Medicine*, 311(13) :812–818. (Cité page 22.)

[217] Zeitouni, B., Boeva, V., Janoueix-Lerosey, I., Loeillet, S., Legoix-Né, P., Nicolas, A., Delattre, O., and Barillot, E. (2010). Svdetect : a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*, 26(15) :1895–1896. (Cité page 65.)

[218] Zhang, J., Zhang, B., Wang, T., and Wang, H. (2015). Lncrna malat1 overexpression is an unfavorable prognostic factor in human cancer : evidence from a meta-analysis. *International journal of clinical and experimental medicine*, 8(4) :5499. (Cité page 146.)

[219] Zhou, J.-D., Yang, L., Zhu, X.-W., Wen, X.-M., Yang, J., Guo, H., Chen, Q., Yao, D.-M., Ma, J.-C., Lin, J., et al. (2015). Clinical significance of up-regulated id1 expression in chinese de novo acute myeloid leukemia. *International Journal of Clinical and Experimental Pathology*, 8(5) :5336. (Cité page 146.)

université PARIS-SACLAY
ÉCOLE DOCTORALE
Cancérologie, biologie, médecine, santé (CBMS)

Title Genetic and epigenetic alterations of chronic myelomonocytic leukemia - Modulation by demethylating agents

Keywords Chronic myelomonocytic leukemia, Genic mutations, Genic expression, DNA methylation, Hypomethylating agents, Bioinformatics

Abstract Chronic myelomonocytic leukemia is a clonal disorder of the hematopoietic stem cell, affecting mainly the elderly. The only curative therapeutic is allogeneic stem cell transplantation, which is rarely feasible. When transplantation is not an option, patients with a severe disease can be treated with a demethylating agent. Thirty to 40% of these patients show hematological improvement, but it remains unknown if these drugs increase overall survival. Analysis of candidate genes by Sanger sequencing, then by New Generation Sequencing, identified about thirty genes that are frequently mutated. These genes encode epigenetic regulators, splicing factors, transcription factors and cell signalling regulators. However, this approach catched only part of the genetic events that characterize this disease. The first objective of this study was to determine the mutational landscape of CMML cells by analyzing the coding and non coding regions of leukemic cell genome. We first performed whole exome sequencing analysis of leukemic and control cells in 49 patients. These analyses showed that in average, a patient carries 14 somatic mutations in its coding regions. We confirmed that the most frequent mutations were in TET2, SRSF2 and ASXL1 genes. We identified also recurrent mutations in 8 new genes, these recurrent mutations occurring at a low frequency. In average, 3 out of the 14 mutations identified in each patient affected recurrently mutated genes. Secondly, we performed whole genome sequencing of leukemic and control cells in 17 patients. These analyses showed that in average, a patient carries 475 somatic mutations in the non repeated regions of the genome. In both the coding and non coding sequences, alterations were observed to be mainly transitions. As a signature of CMML, two mutational processes were identified in all 17 patients and are found in various other cancer types, most likely resulting from the cytosine methylation observed with ageing. A third process, never seen before and without known significance, was also detected in two patients. We collected several samples from 17 patients on a more than two year period : 6 of these patients remained untreated whereas 11 were treated with demethylating agent, among which 6 showed a stable disease and 5 fulfilled criteria of hematological improvement. These sequential analyses showed that 1) the occurence of new mutations is a relatively rare event ; 2) the genetic heterogeneity of the malignant clone is limited ; 3) the mutation allele burden remains unchanged under treatment, whatever the response ; 4) new mutations can appear, even in responding patients. We selected 9 patients, 3 untreated, 3 stable on therapy and 3 responders. We collected monocytes before treatment and a few months later and we analyzed gene expression and DNA methylation in sorted monocytes at these two time points. We did not detect any significant change in gene expression and DNA methylation pattern in untreated patients. In those who responded to treatment, we noticed significant changes in both gene expression, with about 500 deregulated genes, and the DNA methylation pattern, with about 35,000 demethylated regions. In stable patients, the treatment had a limited effect with changes in the expression of about 60 genes, and in the DNA methylation pattern of about 100 regions. These results show that demethylating agents affect gene expression and DNA methylation of responding patients only, suggesting they have mostly an epigenetic effect rather than a cytotoxic one.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Cancérologie, biologie, médecine, santé (CBMS)

Titre Altérations génétiques et épigénétiques dans la leucémie myélomonocytaire chronique - Modulation par les agents déméthylants

Mots-clés Leucémie myélomonocytaire chronique, Mutations géniques, Expression génique, Méthylation de l'ADN, Agents hypométhylants, Analyse bioinformatique

Résumé La leucémie myélomonocytaire chronique (LMMC) est une pathologie clonale de la cellule souche hématopoïétique qui touche principalement les personnes âgées. Le seul traitement curatif de cette maladie est la greffe allogénique de cellules souches hématopoïétiques, souvent difficile à mettre en œuvre. Les patients qui ne peuvent être greffés et dont la maladie présente des critères de gravité se voient proposer un agent déméthylant de l'ADN. Chez 30 à 40% d'entre eux, ce traitement induit une réponse objective dont le bénéfice en terme de survie n'est pas démontré. Le séquençage de gènes candidats a identifié une trentaine de gènes mutés de façon récurrente. Il s'agit de gènes codant des régulateurs épigénétiques, des facteurs d'épissage, des facteurs de transcription, et des protéines de la signalisation intracellulaire. Cette approche ne donnait qu'une vision partielle des événements génétiques associés à la maladie. Le premier objectif de cette thèse a été de recenser l'ensemble des mutations touchant les régions codantes et non codantes de l'ADN dans les cellules leucémiques des patients. Le séquençage de l'exome de cellules malades et de cellules contrôles a été réalisé chez 49 patients. Nos analyses ont montré qu'en moyenne, un patient porte 14 mutations somatiques dans les régions codantes. Nous avons confirmé que les mutations récurrentes les plus fréquentes affectaient les gènes *TET2*, *SRSF2* et *ASXL1*. Nous avons aussi identifié 8 nouveaux gènes mutés de façon récurrente à une faible fréquence. En moyenne, 3 des 14 mutations affectent des gènes touchés de façon récurrente. Le séquençage du génome de cellules malades et de cellules contrôles a été réalisé chez 17 patients. L'analyse réalisée a détecté 475 mutations par patient dans les régions non répétées du génome. Dans l'exome, comme dans le reste du génome, les altérations principales sont des transitions. Deux signatures mutationnelles ont été identifiées et sont observées dans de nombreux cancers, traduisant probablement des altérations de la méthylation des cytosines au cours du vieillissement. Une troisième signature, jamais observée jusqu'alors et de signification indéterminée, a été détectée chez 2 patients. Nous avons alors répété l'analyse de l'exome dans les monocytes triés de 17 patients prélevés de façon séquentielle sur plus de 2 années : 6 n'ont pas été traités et 11 ont été traités par un agent déméthylant, parmi lesquels 6 sont restés stables et 5 ont montré une réponse clinique et biologique objective. L'analyse montre que 1) l'accumulation de mutations est un événement rare ; 2) l'hétérogénéité génétique du clone malade est limitée ; 3) la charge allélique des mutations reste inchangée, même chez les répondeurs ; 4) de nouvelles mutations peuvent apparaître alors que le patient est répondeur. Nous avons alors sélectionné 9 patients, 3 non traités, 3 stables sous traitement sans réponse objective, et 3 répondeurs. Nous avons collecté leurs monocytes avant tout traitement et quelques mois plus tard, alors que 6 d'entre eux étaient traités par un agent déméthylant. Nous avons analysé l'expression des gènes et la méthylation globale de l'ADN à ces deux temps. Chez les patients non traités, nous avons observé une remarquable stabilité de l'expression des gènes et de la méthylation de l'ADN. Chez les patients répondeurs, le traitement induit un changement significatif du niveau d'expression d'environ 500 gènes et la déméthylation d'environ 35000 régions de l'ADN. Chez les patients stables sous traitement, le traitement induit un changement d'expression d'une soixantaine de gènes et du niveau de méthylation d'une centaine de régions seulement. Ces résultats suggèrent que les agents déméthylants n'affectent l'expression des gènes et la méthylation de l'ADN que chez les répondeurs, fournissant un argument important pour un effet essentiellement épigénétique et très peu cytotoxique de ces médicaments.